Evolution of Silica Biomineralizing Plankton

Evolution of Silica Biomineralizing Plankton

A DISSERTATION PRESENTED
BY
BENJAMIN KOTRC
TO
THE DEPARTMENT OF EARTH AND PLANETARY SCIENCES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
EARTH AND PLANETARY SCIENCES

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2013

Dissertation advisor: Professor Andrew H. Knoll Benjamin Kotrc

Evolution of Silica Biomineralizing Plankton

Abstract

The post-Paleozoic history of the silica cycle involves just two groups of marine plankton, radiolarians and diatoms. I apply paleobiological methods to better understand the Cenozoic evolution of both groups. The Cenozoic rise in diatom diversity has long been related to a concurrent decline in radiolarian test silicification. I address evolutionary questions on both sides of this coevolutionary coin: Was the taxonomic diversification of diatoms accompanied by morphological diversification? Is our view of morphological diatom diversification affected by sampling biases? What evolutionary mechanisms underlie the macroevolutionary decline in radiolarian silicification?

Conventionally, diatom diversification describes a steep, monotonic rise, a view recently questioned due to sampling bias. For a different perspective, I constructed a diatom morphospace based on discrete characters, populated through time using an occurrence-level database. Distances between taxa in morphospace and on a molecular phylogeny are not strongly correlated, suggesting that morphospace was explored early in their evolutionary history, followed by relative stasis. I quantified morphospace occupancy through time using several disparity metrics. Metrics describing average separation of taxa show stasis, while metrics describing occupied volume show an increase with time.

Disparity metrics are also subject to sampling biases. Under subsampling, I

find that disparity metrics show varied responses: metrics describing separation of taxa in morphospace are unaffected, while those describing occupied volume lose their clear increases. Disparity can have geographic components, analogous to α and β taxonomic diversity; I find more evidence of stasis in an analysis of $\bar{\alpha}$ disparity. Overall, these results suggest stasis in Cenozoic diatom disparity.

The radiolarian decline in silicification could result from either macroevolutionary processes operating above the species level (punctuated equilibria) or anagenetic changes within lineages. I measured silicification in three phyletic lineages, *Stichocorys*, *Didymocyrtis*, and *Centrobotrys*, from four tropical Pacific DSDP sites. Likelihood-based model fitting finds no strong support for directional evolution, pointing toward selection among species, rather than within species. Each lineage shows a different trajectory, perhaps due to differences in the ecological role played by the test. Because *Stichocorys* shows close correspondence to the assemblage-level trend, abundance may be an important factor through which within-lineage changes can influence the macroevolutionary pattern.

Contents

Listing of figures

Meinen Grosseltern, Fritz und Trude

For my grandparents, Fritz and Trude

Acknowledgments

"If I have seen further", Newton famously wrote, "it is by standing on ye shoulders of giants." While I hope that I have adequately acknowledged the relevant giants in the chapters that follow, I also want to thank the veritable army of supporters who have hoisted, pulled, and cajoled me onto my own giant-shoulder perch.

My advisor, Andy Knoll, opened up this opportunity for scholarship in an email exchange in 2005, and has not ceased to provide opportunities and support since. His unflagging trust, his willingness to treat a student as a peer from the outset, and his encouragement to have the courage of my convictions buoyed me when I felt dwarfed by many academic challenges. I thank Charles Marshall for his gifted paleobiological metaphors. While serving on my committee, he brought a rigorous intellectual engagement to my work at its earliest stages that remained unparalleled throughout my graduate school experience. Jacques Dumais, also a past committee member, was generous with project ideas, conversations, and with the use of his lab and microscopy equipment. I thank Dave Johnston and Jerry Mitrovica for joining my committee after Charles and Jacques left Harvard. What better way to express my gratitude than in Jerry's own words? "I would like to thank, first and foremost, Jerry Mitrovica for his insight, wit and generosity. When I think of greatness, I think of Jerry." Inside jokes aside, I want to thank him for the indirect lesson that personal warmth and kindness are legitimate factors in choosing a scientific collaborator, and for the direct lesson

that there are academics out there who fit the bill.

I am grateful for the generous financial support I received. The EPS department footed the bill for my first year in graduate school, provided a travel grant for museum work in Berlin, and funded memorable field experiences in Arizona and Hawai'i. I thank the Agouron Institute for two field courses, in Oman and the western US. I also owe my gratitude to the late Alfred W. Stickney for endowing a fellowship that supported me during my second year at Harvard, and to the PlanktonTech Helmholtz Virtual Institute for financial support during subsequent years.

I am indebted to Dave Lazarus for his help at the Museum für Naturkunde in Berlin. Chapter 5 would not have been possible without his generosity in teaching me the necessary taxonomy and providing unfettered access to the MRC collection haystack—along with the needle-finding expertise to locate workable samples within. Brian Huber granted access to the equivalent haystack at the Smithsonian.

My thanks are also due to those scientists who shared their data freely for use in the work presented here: Dan Rabosky and John Alroy provided their code for subsampling analyses, while Ulf Sörhannus shared his sequence alignments and phylogenetic trees.

On the large number of my PhD projects that never reached fruition, I received help from an even larger number of people. Zoe Finkel hosted me at Mt Allison University while giving me a crash course in diatom culture, provided samples for a FIB-SEM study of fossil diatoms, and subsequently provided the opportunity for collaboration on a paper. Annika Sanfilippo shared copious notes and data on Cenozoic radiolarian lineages, Missy Holbrook and Colleen Cavanaugh gave me access to their lab resources for projects investigating the mechanical strength of diatom frustules, Jan Michels taught me how to image diatom frustules by CLSM during a short visit at the Universität Kiel, Tanja Bosak provided lab space at MIT for diatom culturing, and Sébastien Besson collaborated on a generative morphospace for diatom outlines.

I benefitted considerably from both the tutelage and the camaraderie of my

fellow Knollites. Postdocs Sara Pruss, Dave Johnston, Nick Tosca, Ben Gill, and Tais Dahl were shining examples to look up to—of not just good scientists, but good people. Rowan Martindale will additionally be remembered for proffering chocolate at all the right moments. Walton Green not only contributed many helpful paleobiological conversations to my development, but also stepped in as *R* tutor, statistical consultant, proofreader, wedding officiant, and true friend. I feel exceptionally lucky, professionally and personally, that our paths crossed. My academic 'older siblings' provided much-needed guidance and support in the first years of graduate school; I particularly appreciated Tony Rockwell's wisdom, Phoebe Cohen's friendship and empathetic support through the rougher patches, and Jon Wilson and Woody Fischer for patiently explaining carbon isotope fractionation (among many other fundamentals) over and over again (not to mention the dadaist skits for St Barbara's day). I particularly wish to thank JC Creveling. I was helpless in advising her when she first arrived as the younger graduate student; by the time she left, she was the sage confidante I turned to for advice, and I am proud to call her an 'older sibling', colleague, and friend.

Outside the Knoll lab, Allie Gale was an invaluable ally in our first-year battle with math; Allison Shultz helped me wrangle molecular data in *R* many years later. David Hewitt helped pass many an hour with distracting visits, ripe with scientific and entertainment value alike. I must also acknowledge the tirelessly supportive efforts of the EPS administration, particularly Chenoweth Moffat, Sarah Colgan, Cindy Marsh, Maryorie Grande, Bridget Mastandrea, and Paul Kelley. I thank Liliana Umana for keeping my office spotless and my Spanish from going down the drain.

To Beaudry Kock, I owe a debt of gratitude I cannot describe. Thank you for what must have been well over three hundred weekly meetings of *Doctoral Students Anonymous*, for asking the tough questions and remembering to ask for the deliverables, for praise and encouragement when I most needed it, and for really understanding what it's like; for proofreading, paired programming, and above all, friendship. This thesis would literally not have been possible without you.

I would like to thank the exceptional circle of local friends who provided much-needed support, distraction, and perspective outside the tunnel vision of academia: Evan and Katie, Monique, Mark and Mattie, Pierre and Nicole, Audrey, Sarah Jane and Will, Travis and Marcy, Nick and Maggie, Jenny and John, and Mateo and Vanessa. I am humbled by the love, support, and joy my family in Florida has given me; breaks with Hugh, Nancy, Summer, Gabe, and Anna were truly restorative.

I have had the great fortune of inspiring teachers, among whom I hold Colin Strange and Peter Allison responsible for sparking my interest in geology and paleontology, respectively; without them, I wouldn't have made it here. I am profoundly grateful to my parents, Peter and Maria, who paved the way for my success when they instilled in me a love of learning, and encouraged me to take the opportunities I was so lucky to be afforded—even when it meant moving half a world away. Last, and the opposite of least, I want to express my deepest gratitude to my amazing wife, Kati, who supported, fed, and loved me, even when I was not at my best. Thank you for giving me the reason to stay and the strength to follow through.

The researches of many commentators have already thrown much darkness on this subject, and it is probable that, if they continue, we shall soon know nothing at all about it.

Mark Twain

1

Introduction

The emerging discipline of geobiology seeks to understand the interactions between Earth and life in deep time. Biomineralizing organisms are of particular interest to geobiologists because their evolution is both tied to the geochemical cycling of the material from which their skeletons are made, and capable of modifying that geochemical cycle. Importantly, biomineralization also renders organisms susceptible to preservation in the fossil record, making them accessible for paleobiological study. This is especially true of marine planktonic microfossils, which can accumulate in deep sea sediments at high temporal resolution and in such great numbers that sample sizes are limited only by the effort available to analyze them.

The silica cycle is an attractive system for geobiological inquiry because it is decoupled from the complexities of Earth surface redox chemistry and has been dominated by a small cast of biological actors. The predominant narrative in the

post-Paleozoic history of the marine silica cycle involves just two groups of marine plankton, the radiolarians and the diatoms, unicellular clades with an extensive fossil record. In this dissertation, I apply paleobiological methods to better understand the Cenozoic evolution of both of these groups.

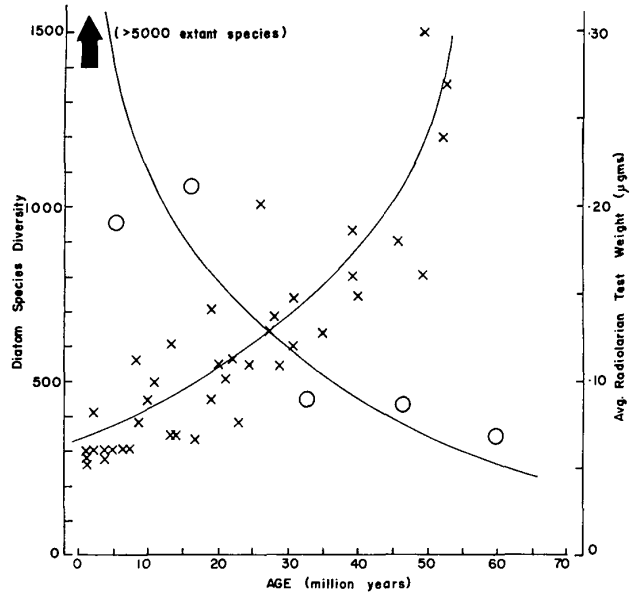


Figure 1.1: Figure reproduced from Harper and Knoll (1975), showing a Cenozoic rise in diatom diversity (circles) accompanied by a decline in radiolarian silicification, measured by test weight (crosses).

The evolution of diatoms and radiolarians is interwoven by their common physiological requirement for silica. The observations linking their evolutionary histories were first assembled into a coherent coevolutionary narrative by Harper and Knoll (1975), who related a Cenozoic rise in diatom diversity to a concurrent decline in radiolarian test weight (Fig. 1.1). They suggested that as diatoms rose to ecological prominence, silica concentrations in the surface oceans declined, eliciting an evolutionary response towards less silicified tests in radiolarians. In this dissertation I address evolutionary questions arising on both sides of this coevolutionary coin. Was the Cenozoic taxonomic diversification of diatoms

accompanied by morphological diversification? To what extent is our view of diatom diversification affected by sampling biases? And what evolutionary mechanisms underlie the macroevolutionary decline in radiolarian silicification?

While the data available to document Cenozoic diatom diversification have improved significantly since Figure 1.1 was plotted (Small 1946), the conventional view of diatom diversification has remained similar, describing a sharp, monotonic rise (Spencer-Cervato 1999). This view, however, was recently called into question by Rabosky and Sorhannus (2009), who applied a variety of subsampling methods to account for temporal differences in sampling intensity. The subsampled diversity curves, in stark contrast, show peak diversity around the Eocene/Oligocene boundary, followed by a steep decline in diversity and a gradual Neogene recovery that never reaches the highest Paleogene levels.

In Chapters 2 and 3, I open a different window on diatom diversification in light of these controversial results: the record of morphological disparity. In Chapter 2 I describe the construction of a diatom morphospace based on discrete morphological characters and populated through time using an occurrence-level database, *Neptune*. I compare this morphospace to a molecular phylogeny of diatoms and find little correspondence between the two, which I interpret to imply that morphospace was explored early in their evolutionary history, followed by relative stasis. I then describe the occupancy of the morphospace using several metrics of disparity, and find that metrics describing the average separation of taxa in morphospace (i.e. how similar taxa are to one another) also show stasis. Metrics describing the volume of occupied morphospace (i.e. the total range of explored morphologies), however, show an increase with time.

In Chapter 3, I confront the possibility that disparity metrics, much like measures of taxonomic diversity, can suffer from sampling biases, because both derive from the same fossil record. Since the diatom morphospace described in Chapter 2 was populated with a database of fossil occurrences, I apply methods of subsampling established for diversity studies to this diatom morphospace. Consistent with earlier studies (e.g. Ciampaglio et al. 2001), I find different responses to subsampling among the different disparity metrics. Those metrics

describing the separation of taxa in morphospace appear immune to changes in sampling, while those describing the volume of occupied morphospace no longer show clear increases under subsampling. I introduce the notion of geographic components in disparity, analogous to α and β taxonomic diversity, and find more evidence of stasis in an analysis of $\bar{\alpha}$ disparity. Overall, the results suggest that the Cenozoic history of diatom disparity was broadly characterized by stasis.

Chapters 2 and 3 are formatted for submission to the journal *Paleobiology*, and are intended for joint publication in the same issue.

In Chapter 4, I review the results of the preceding chapters and place them in context through a review of diatom diversity studies and subsampling methods. This chapter was submitted to the edited volume *Evolution of Lightweight Structures*, which emerged from a collaboration with the PlanktonTech Helmholtz Virtual Institute and is to be published by Springer Verlag.

Finally, in Chapter 5, I turn to the radiolarian response and investigate the mechanisms underlying the assemblage-level decline in silicification. This decline could, in principle, be due to purely macroevolutionary processes operating above the species level, as predicted by the strictest formulations of punctuated equilibria (Eldredge and Gould 1972), but it could also result from anagenetic changes within species. I made detailed morphometric measurements of silicification on three well-defined anagenetic lineages. Using likelihood-based model fitting (Hunt 2006), I find no strong support for directional change, suggesting that selection among species, rather than within species, is at play. Each lineage shows a different trajectory, which may be due to differences in the autecological role played by the test, and its relationship to feeding strategy. One lineage, however, shows close correspondence to the assemblage-level trend, suggesting that species abundance may be an important factor, and that within-lineage changes can in fact influence the macroevolutionary pattern.

Together, these studies paint a more refined picture of the story told by Figure 1.1. They highlight the importance of accounting for geological biases when examining macroevolutionary trends, and provide a novel path to do so through the window that morphospace studies provide on the diversification

history of clades. In the case of diatoms, they suggest that most of the range of morphologies may have been in place at the outset of the Cenozoic. They suggest that the selective pressures on radiolarians may not have come from a gradual increase in diatom diversity, but from sharp shifts in the Paleogene, or from changes in the silica cycle due to changes in abundance or preservation. Finally, they provide an insight into the complex trajectories of individual lineages underlying macroevolutionary patterns. These results show neither the immutable stasis expected under strict punctuated equilibria, nor the directional change expected under anagenetic, within-species selection. They are a reminder that even when obscured by the abstraction of a macroevolutionary compilation, each and every species—including lowly oceanic microbes—is a complex biological entity subject to interactions as intricate as the beautifully ornate shells they leave behind.

Es scheint mir bei der Wichtigkeit, welche die Natur selbst diesen kleinen Organismen ertheilt, die sie zwar in individueller Energie weit unter Löwen und Elephanten, in ihrem allgemeineren socialen Einflusse aber weit über dieselben gestellt hat… nicht unangemessen, einige neuere Beobachtungen … anzuschließen.

Christian Gottfried Ehrenberg

2

Disparity stasis in a morphospace of planktonic marine diatoms

Abstract

Both molecular clocks and the first appearances of major groups in the fossil record suggest that most of the range of diatom morphologies had evolved by the end of the Cretaceous. A canonical reading of the Cenozoic fossil record, however, suggests a dramatic rise in taxonomic diversity during the Cenozoic Era that can be interpreted as an explosion of morphological variety. We investigated this apparent discrepancy using a discrete-character-based, empirical diatom morphospace, resolved by molecular phylogeny and by fossil occurrences through time. The morphospace shows little correspondence to phylogeny and little Cenozoic change in disparity as measured by mean pairwise distance. There is, however, an increase in the total volume

of morphospace occupied. Although the increase in occupied volume through time superficially supports a conclusion of increasing morphological variety, sampling biases and other data suggest an underlying stasis, which is more consistent with the molecular clock data.

2.1 Introduction

Diatoms are a diverse and ecologically important part of the autotrophic ocean plankton, responsible for a substantial proportion of total photosynthesis globally (around 10–20%, according to estimates of Raven, 2003, and Nelson, 1995). Beyond their importance at the base of the food web, diatoms are particularly important to the global carbon cycle because they sink readily and thus export carbon from the surface ocean (Dugdale and Wilkerson 1998). This is due in part to their relatively large cell size and growth in chains and blooms, but also to the ballast provided by their silicified cell walls, or frustules.

Diatom frustules are highly preservable and can accumulate in great numbers in marine sediments, giving marine planktonic diatoms an extensive fossil record that stretches back at least to the early Cretaceous Period. Their abundance and morphological diversity makes them useful as biostratigraphic markers, particularly in the Cenozoic Era, and thus extensive data exist about their occurrence through time. The *Neptune* database (Lazarus 1994; Spencer-Cervato 1999), for example, is a compilation of tens of thousands of records of diatom occurrences in deep sea sediment cores drilled by the Deep Sea and Ocean Drilling Programs (DSDP and ODP). This represents the combined output of many decades of micropaleontological effort and provides a rich and readily available data set for macroevolutionary studies.

Among the macroevolutionary questions that have been addressed concerning diatom fossils—including their diversity history (Spencer-Cervato 1999), biostratigraphy (Fenner 1985; Barron 1985), coevolution with cetaceans (Marx and Uhen 2010), and the Cenozoic silica cycle (Harper and Knoll 1975; Lazarus et al. 2009)—the relationship between their taxonomic and morphological

diversification stands out as unresolved.

Because fossil taxa are defined morphologically, the number of distinct taxa is, by definition, a measure of morphological variety. But this variety can also be measured with much more nuance by quantifying aspects of shape directly, then summarizing these measurements by a variety of metrics of disparity (Erwin 2007). Both diversity and disparity are used in macroevolutionary studies of groups with an extensive fossil record, including the biomineralizing microplankton. The two measures provide different views of evolutionary change through time, and do not necessarily vary together.

On the contrary, many examples of decoupled changes in diversity and disparity have been documented; clades frequently follow a pattern of rapid filling of morphological space at low taxonomic diversity early in their history (reviewed in Foote 1997, p. 137). This pattern has been referred to as "asymmetric diversification" (Webster 2007). Perhaps the most famous large-scale example is the Cambrian explosion, when the major animal body plans evolved early (high disparity), leaving the rest of the Phanerozoic to play out in relative macromorphological stasis while taxonomic diversity increased (e.g. Gould 1989; Erwin et al. 2011). In this study, we examine whether this pattern is also common to the diatoms.

The history of diatom taxonomic diversity has been conventionally taken to support major morphological diversification late in the group's history, associated with a steep rise to ecological prominence through the Cenozoic Era. Other lines of evidence, however, suggest that diatoms may have experienced relatively broad morphological stasis over the past 65 million years: both molecular clocks (Kooistra and Medlin 1996; Sorhannus 2007) and fossil discoveries (reviewed in Sims et al. 2006) suggest that all major morphological groups of diatoms were present by the end of the Paleocene Epoch. The question of whether the suggested Cenozoic evolutionary history of the diatoms is better described in terms of diversification or stasis has become increasingly intriguing with recent work suggesting that the Cenozoic rise in taxonomic diversity may largely be an artifact of sampling bias (Rabosky and Sorhannus 2009). This makes clear the

need for a different—morphological—window on the Cenozoic evolutionary history of the diatoms. In this study, we review the evidence for both diversification and stasis, and use a morphospace to gain a more differentiated view of Cenozoic diatom evolution.

2.2 Diatom Diversity and Disparity

2.2.1 The Importance of Frustule Shape

The shape of the diatom frustule is ecologically and thus evolutionarily important because the frustule performs a variety of functions. Indeed, the frustule has been implicated as a key innovation allowing the diatoms to rise to their present-day ecological importance (Kooistra et al. 2007; Hamm and Smetacek 2007). While the diatom frustule has not been definitively shown to perform any one single function to the exclusion of all others, a number of hypotheses have been presented, which can be summarized under two major headings: those based on a top-down view of diatom evolution, driven by predation, and those based on a bottom-up view, driven by resource competition.

The top-down view sees the frustule as a way to decrease mortality, providing defense against the crushing mouthparts of grazers through mechanical strength and deterrent spines (Smetacek 2001; Hamm et al. 2003), and a rigid barrier against the entry of pathogens or parasites (Smetacek 1999). The ballast provided by frustules may also facilitate the sinking of infected cells from surface populations (Raven and Waite 2004). In contrast, the bottom-up view sees the frustule as a key to the diatoms' ability to take up nutrients rapidly and store them over several generations by providing ballast to counteract the buoyancy of the vacuole and rigidity against its turgor pressure (Raven and Waite 2004), as well as allowing cells to sink out of depleted surface waters to nutrient-enriched depths (Raven 1997; Raven and Waite 2004).

Since biological structures generally can perform a variety of functions (Dudley and Gans 1991; Marshall 2003), it is quite feasible that the frustule

confers a fitness advantage in multiple ways. Ascribing multiple functions to the diatom frustule might also help explain the diversity of forms seen, since adding design constraints theoretically increases the number of equally "fit" solutions, that is, morphologies (Marshall 2003; Niklas 2004). Thus, given its importance to both top-down and bottom-up drivers of evolution, frustule morphology is centrally important to any understanding of diatom evolution.

2.2.2 History of the Major Diatom Groups

Mesozoic Origins

Diatoms have been divided into four major taxonomic groups characterized by different gross morphological types: forms with round (1), multi-angled (2), or bilaterally symmetrical (3) outlines, and slit-bearing (4) forms. The frustule of radial centric diatoms (1) have a ring-shaped structural "pattern center" (an imperforate siliceous structure from which the ribs giving rise to the rest of the frustule originate during morphogenesis). The two valves making up the frustule are generally circular in plan view, i.e. they are radially symmetrical. The bi- and multipolar centrics (2) share the same ring-shaped pattern center, but have valves that are often elongated and distorted in plan view, often with well-delimited areas of smaller pores that seem to be involved in mucilage secretion. The pennate diatoms (3) are characterized by a linear pattern center and generally have a bilaterally symmetrical (pennate meaning feather-shaped) valves. The raphid diatoms (4), a subgroup of the pennates, possess a slit in the surface of the valve through which part of the protoplasm can be extruded for locomotion over substrates.

Molecular phylogenies of diatoms broadly agree on an order of divergence for these four major groups (Medlin and Kaczmarska 2004; Damsté et al. 2004; Sorhannus 2004). The raphid pennates appear to form a monophyletic group, and while the radial centrics form a clade in some treatments, the bi- and multipolar centrics and the araphid pennates are generally considered to be paraphyletic. While they differ in many details, published molecular phylogenies

of diatoms all show the radial centric diatoms as the most basal group, from within which diverge, or which are sister to, the bi- and multipolar centrics. Pennate diatoms are nested within the bi- and multipolar centrics, with raphid pennates forming a derived clade within the pennates. These relationships predict an order of first appearances for these four groups that is confirmed by the fossil record (Sims et al. 2006).

Beyond order of appearance, the most recent molecular-clock estimates of divergence times (Sorhannus 2007) suggest that all four major groups appeared in the Mesozoic. Actual first appearances based on fossils postdate these estimates by 10–40 Myr. Since the fossil record is imperfect (and also, to a lesser extent, because of the lag between coalescence and speciation), a lag between molecular divergence and first appearance in the fossil record is expected. The oldest fossil diatom occurrence accepted by Sims et al. (2006) is a radial centric from the Jurassic of Germany (Rothpletz 1896), roughly the same age as that predicted by Sorhannus' molecular clock (though more recent discoveries have suggested that the earliest fossil diatoms may be from non-marine environments, Harwood et al. 2007). The molecular divergence time for bi- and multipolar centrics is 150 Ma, around 40 Myr before their Aptian-Albian first appearance in deposits of the Weddell Sea (Gersonde and Harwood 1990). The first pennate diatoms appear in the fossil record of the Campanian (Sims et al. 2006), some 40 Myr after their predicted divergence. The molecular clock predicts the divergence of raphid pennates at 74 Ma; their first appearance in the fossil record follows about 10 Myr later in the Paleocene of Russia (Pantocsek 1886; Witt 1886). The magnitude of these differences between the molecular and fossil estimates of first appearance is comparable to other groups (Sperling et al. 2011, for example, cite around 20 Myr for early brachiopods), particularly considering that open ocean habitats may encourage longer gaps between speciation and first appearance in the fossil record (Anderson et al. 2011). Crucially, however, both molecular clocks and the fossil record indicate that the four major groups (and thus highest-level taxa) of diatoms had evolved by the earliest Cenozoic Era.

Given the largely Mesozoic origin of the four major diatom taxa, and the gross

morphotypes they represent, we might expect relative stasis in morphospace occupation through Cenozoic time.

Cenozoic Events

Following the Mesozoic establishment of the four major groups, the molecular and fossil records show three major Cenozoic events in diatom evolution, according to Sims et al. (2006) and Kooistra et al. (2007). These major events are (1) the invasion of fresh water, (2) the evolution of the Thalassiosirales (a subgroup of the bi- and multipolar centric diatoms with a round outline that is ecologically important in modern oceans), and (3) the evolution of the raphe. While the invasion of fresh water obviously represents an important event for the clade, since the bulk of its present-day diversity is found there, it is difficult to see how this would influence morphological diversity in the open ocean. Furthermore, fossil evidence from terrestrial deposits of Early Cretaceous age in Korea (Chang et al. 2003) and Late Cretaceous age in Mexico (Chacón-Baca et al. 2002) suggest that terrestrialization may have begun earlier than commonly thought. The evolution of the raphe, and hence the diversification of raphid diatoms, is probably phylogenetically the most important event of the Cenozoic Era, because there is so much diversity in that group.

2.2.3 Cenozoic Taxonomic Diversity

The Cenozoic stasis in planktonic diatom morphospace suggested by molecular clocks stands in stark contrast to a canonical reading of the Cenozoic record of planktonic diatom diversity. The record of diatom species diversity has long been interpreted as an almost monotonic increase through the Cenozoic Era (Small 1946; Spencer-Cervato 1999), though this view has been recently challenged by Rabosky and Sorhannus (2009). This canonical view has been widely accepted and marshalled as evidence, for example, in explanations of Cenozoic decline in marine silicic acid concentrations (Harper and Knoll 1975; Lazarus et al. 2009) and the evolution of modern phytoplankton (Falkowski et al. 2004). Such

explanations of Cenozoic diatom evolution imply that their sharp rise in diversity is a proxy for dramatic environmental expansion and success. In so far as ecology and morphology are linked, it would be reasonable to expect that ecological diversification would go hand in hand with an increased diversity of form. The canonical reading of the diatom diversity record, therefore, implies a major ecological expansion of the diatoms in Cenozoic Era, and, if not directly requiring an expansion of morphospace, certainly suggests it.

In this study we test the hypothesis that, in spite of increasing taxonomic diversity, disparity and morphospace occupancy of marine planktonic diatoms through the Cenozoic Era were characterized by stasis. Prior morphospace studies on diatoms, including both theoretical (Pappas 2005) and empirical (Du Buf and Bayer 2002) morphospaces, were limited either to particular lineages or studies of valve outlines and pennate striations, ignoring the many other features of frustules. Because of the diversity and complexity of structures comprising the diatom frustule, we opt to describe diatom morphology using discrete characters (on the nominal scale of Stevens 1946). We use the record of diatom occurrences provided by the *Neptune* database to quantify occupancy of this morphospace through time. In order to cover the full breadth of morphologies captured by this record, we work at the genus level and use the diatom genera found in the *Neptune* database to construct a morphospace. We first discuss ways of visualizing morphospace to depict more explicitly the morphological meaning of morphospace ordinations. With this more intuitive sense, we interpret the history of Cenozoic diatom disparity.

2.3 Materials and Methods

The full range of morphologies a group of organisms can have is often described as a morphospace. Morphospaces are vector spaces defined by axes representing an aspect or measurement of the organism. Each point in these spaces represents a distinct morphology, which may or may not be occupied by an organism. A distinction is commonly made between theoretical morphospaces and empirical

morphospaces. In *theoretical* (or generative) morphospaces, the axes are the parameters of a geometric model of organism shape. Theoretical morphospaces tend to be of relatively low dimensionality, and are thus often graphically represented in their full dimensionality; the classical and foundational example is Raup's three-dimensional morphospace of coiled shells (Raup and Michelson 1965). In *empirical* morphospaces, by contrast, each axis represents a measurement of some sort, which could be a continuous measurement like a length or an angle, or a discrete measurement such as the number of segments or the presence/absence of a feature. Empirical morphospaces tend to be high-dimensional, and thus require projection or ordination to be visualized in two dimensions. While there has been some debate about the relative merits of theoretical versus empirical morphospaces (e.g. McGhee 1999), both can be considered as different manifolds within a "true" phenotypic morphospace comprised of more dimensions than can either be modeled or measured. Depending on the organisms and the research questions at hand, either a theoretical or an empirical morphospace may be the most relevant representation of the range of morphologies occupied by a group of organisms.

2.3.1 The *Neptune* Database

Documenting the occupation of morphospace through time requires measures of a taxon's morphology as well as its range in time. In many morphospace studies to date, the latter has been achieved through range compilations, inferring a taxon's duration based on first and last occurrences (e.g. Foote 1993, 1995a; Smith and Bunje 1999; Eble 2000a). Over the past two decades, however, paleobiologists have begun to assemble and use large databases of fossil occurrences so as to address secular differences in sampling. In this study we thus use an occurrence-based database to populate a morphospace through time.

The *Neptune* database provides a record of Cenozoic planktonic diatom occurrences. Sampling intensity in *Neptune* is not uniform through time: the number of samples decreases substantially with age, in part because older seafloor

is more likely to have been subducted. Because more recent sediments are found almost everywhere on the ocean floor, any drilling operation to older sediments will also penetrate younger sediments, inflating the number of younger samples.

We constructed a morphospace using discrete characters, populating it through time using the occurrence data from the *Neptune* database. We coded 123 discrete morphological characters for 152 diatom genera using descriptions from the taxonomic literature. These genera represent all the valid genera found in the *Neptune* database (Lazarus 1994; Spencer-Cervato 1999), plus those found in the three published Cretaceous diatom assemblages recovered by the DSDP/ODP program (Hajós 1976; Gersonde and Harwood 1990; Fourtanier 1991). Genera described as resting stages, which represent a non-vegetative stage of the life cycle and sometimes radically different morphologies, were excluded from the analysis. By linking these morphological data with the fossil occurrence data in the *Neptune* database, we were able to reconstruct diatom morphospace through time in the open ocean. Over 95% of the diatom occurrences in the *Neptune* database are from cores drilled at water depths $>$ 1000 m (and 70% from depths $>$ 2000 m); thus, the evolution of diatoms in coastal or terrestrial environments may have followed quite different trajectories.

2.3.2 Choice of Characters

We compiled a list of morphological characters from general descriptions of frustule morphology (Barber and Haworth 1981; Anonymous 1975) and taxonomic descriptions of the chosen genera. To avoid introducing bias from the taxonomic structure inherent in commonly used terminology, we formulated morphological characters as generally as possible.

For many aspects of diatom morphology, the same shape or structure is given different names in the literature depending on the taxonomic group in which it appears. For example, some authors use almost non-overlapping vocabularies in describing pores and their arrangement on the frustule in the two major groups of diatoms, centrics and pennates, although the structures are obviously

comparable (see, for example, Anonymous 1975). Since coding separate characters for "areolation" (p. 348, ibid.) vs. "striation" (p. 349, ibid.) would introduce an artificial separation between similar structures, we instead created generally applicable characters for "pore arrangement". This single set of characters can represent the morphologies bearing different sets of names in the two groups. We applied a similar, taxonomically agnostic approach to other cases where the terminology used in the literature for similar structures differs among genera because the structures differ developmentally, are not considered homologous, or simply occur in different taxa.

The characters chosen in this way were coded as binary or unordered multistate characters (i.e. they are measurements on the nominal scale, Stevens 1946). Although all missing data were treated equally in the analysis presented below, we distinguished between three different types in the morphological data matrix: character states not observed because of missing information, logically inapplicable character states, and character states varying within or between species of a genus with no obviously predominant state. A description of each character is provided in Appendix B on page 187, while the complete morphological data matrix is provided in Appendix C on page 194.

2.3.3 Morphological Data

We coded the morphological character states for each genus based on descriptions from the taxonomic literature. For 64 of the 152 genera investigated, we used descriptions provided in the standard text by Round et al. (1990). For the remaining genera, we consulted the wider literature, usually the original generic description as well as the most detailed or recent study available, and sought SEM images wherever possible. A complete listing of the sources consulted for each genus is provided in Appendix D on page 230.

Because of the sources of incomplete data mentioned above, some of the genera in the data matrix had relatively few characters with valid states. Likewise, a number of the characters had valid states for only a few genera. In order to avoid

including relatively uninformative genera and characters, we removed genera and characters with less than 80% observed entries. The implications of setting data culling thresholds have been discussed by Ciampaglio et al. (2001) and are investigated in detail in Chapter 3. The culled data matrix consists of 140 genera and 100 characters (Appendix C).

2.3.4 Occurrence Data

Diatom occurrence data, used in the analysis to determine how the morphospace became occupied through time, were downloaded from the *Neptune* database via http://portal.chronos.org/ in May, 2009. A substantial number of changes were made, including correcting misspelled genus names, eliminating zero-age occurrences, eliminating taxa incorrectly classified as diatoms, and eliminating taxa considered to be resting stages rather than vegetative cells (according to Hargraves 1986; Harwood 1988; Hendey and Simonsen 1972; Suto 2004, 2005; Suto et al. 2009, 2011). Because the *Neptune* database only contains microfossil occurrences from the Cenozoic Era, compound taxon lists from the three described Cretaceous DSDP/ODP assemblages were added to the occurrence dataset (Hajós and Stradner 1975; Gersonde and Harwood 1990; Fourtanier 1991).

2.3.5 Software

The analyses described below were carried out using the statistical programming language *R* (R Development Core Team 2011). The code needed to run the analyses, as well as the plotting software, is provided in Appendix F on page 239; all files are also provided in the online supplement.

2.4 Analysis

2.4.1 Low-Dimensional Representation of the Morphospace

Principal coordinates analysis (PCO) is one of several ordination methods used to represent high-dimensional data sets in low dimensions. Such methods are needed to plot the 100-dimensional, nominal-scale morphospace (consisting of discrete, unordered characters) defined by the morphological data matrix in two or three continuous dimensions.

In a better-known ordination method, principal components analysis (PCA), an $m \times n$ data matrix is transformed directly (where m is the number of genera and n is the number of characters). In contrast, the algorithm for PCO (Gower 1966) operates on an $m \times m$ matrix of pairwise dissimilarities between taxa. When these dissimilarities are distances of the familiar Euclidean sort, PCO produces an equivalent result to PCA. In the present case, however, the genera reside in a space defined by discrete, unordered states, so a different metric of dissimilarity is required. We used the sum of character state mismatches divided by the number of possible matches (i.e. excluding comparisons with invalid character states) as the measure of dissimilarity, also used, for example, by Foote (1999), Lupia (1999), and Boyce and Knoll (2002). This dissimilarity metric has the advantage that it accounts for similarity where a valid comparison can be made, but does not inflate dissimilarity by scoring mismatched states where one taxon has invalid or inapplicable states.

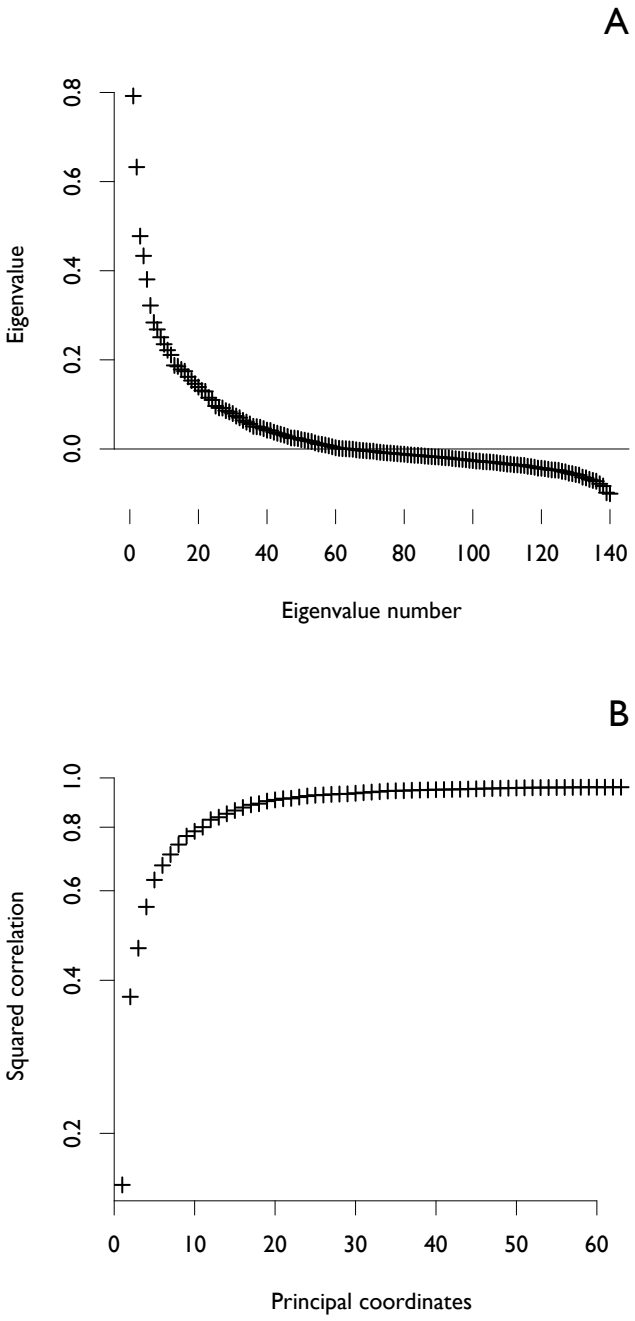
Variance Explained by PCO Axes

Before interpreting the distribution of taxa in a two-dimensional morphospace ordination, it makes sense to consider how well the first two axes represent the full space. There are two basic approaches to calculating this; once can either compare the eigenvalues associated with PCO axes or correlate distances in PCO-space with original distances. The methods give slightly different results.

Since PCO is an eigenvector method, a natural first approach is to compare the

Figure 2.1 ***(following page)*****:** Plots showing the distribution of variance among the principal coordinate axes. A, the magnitude of eigenvalues associated with the PCO axes, which is indicative of their relative information content. Although the higher eigenvalues account for much of the total, suggesting that much of the information is contained in them, the first two PCO axes do have much larger associated eigenvalues, and the inclusion of further axes shows rapidly diminishing returns. B, the squared correlation (R^2) between squared pairwise dissimilarities in the original $(m \times m)$ matrix and squared Euclidean distances in a PCO-space (y-axis) including increasing numbers of PCO axes (x-axis).

Figure 2.1: (continued)



eigenvalues associated with first two principal coordinate axes to those associated with the higher axes. Plotting those values (Fig. 2.1A) provides a qualitative assessment of the variance associated with each axis, showing that the eigenvalues drop rapidly, although the higher axes are not negligible. One way to quantify this is to divide the sum of the first two eigenvalues by the sum of all eigenvalues (as done by Boyce and Knoll 2002; Foote 1995a), giving an estimate of 26% of the total variance explained by the first two principal coordinate axes. However, only 63 of the 140 eigenvalues are positive (see Fig. 2.1A). This could be due to several reasons: first, we should not expect more positive eigenvalues than characters; second, there were missing data; and finally, because the dissimilarity metric chosen is non-Euclidean, there may not be an arrangement in the (Euclidean) PCO-space that corresponds to the calculated dissimilarities.

There are several ways to deal with these negative eigenvalues in estimating the information from the original data matrix in the principal coordinate axes. The `cmdscale()` function that carries out PCO in *R*, for example, calculates a "goodness of fit" statistic in two ways that are both different from the above: either negative eigenvalues are ignored, which results in the estimate of variance explained dropping to 18%, or the sum of the absolute values of the eigenvalues is used instead, in which case the estimate drops even further to 14%.

An empirical alternative for estimating the information retained by the principal coordinate axes is to calculate the correlation between pairwise distances among genera in the original dissimilarity matrix and the pairwise distances of the same genera in PCO-space (Foote 1999). As expected, including progressively more principal coordinate axes increases the correlation (Fig. 2.1B). This approach suggests that the first two principal coordinate axes explain about 37% of the variance in the original dissimilarity matrix, a higher value than the estimates based on comparing eigenvalues.

It is also useful to know which characters contribute most to each of the PCO axes. While it not possible to plot "loadings" (the projection of the the original character axes into the lower-dimensional space), as commonly done for PCA, because our characters are discrete, unordered, and contain missing data, Foote

(1995b; 1999) suggested an analogous approach to discover which characters are associated with which PCO axis. The idea is to compare the character states of taxa for each character with the PCO scores of taxa using a nonparametric measure of correlation. One such measure is the Cramér coefficient, which can be used to measure the degree of association between attributes which are measured in unordered categories (Siegel and Castellan Jr. 1988, p. 225). We calculated this measure for each pairing of characters and PCO axes. In order to discretize the PCO scores, we divided each axis into four arbitrary intervals of equal length. We then constructed a $j \times 4$ contingency table, where j is the number of valid character states for the character in question. Entries in the table are counts of the number of genera, for example, with character state 0 and falling in the lowest quarter of the range of the PCO axis. Measuring an association between score on the PCO axis and character state requires at least two columns in this contingency table to have nonzero sums, which is why characters that had fewer than two states with valid entries were culled from the dataset. From this table, we calculated a Cramér coefficient and an associated p-value using the `assocstats()` function in the *R* package *vcd* (Meyer et al. 2011). The results of the 6426 pairwise comparisons are summarized in Figure 2.2.

While the associations between morphological characters and PCO axes are strongest in the lower axes, there are also significant associations with higher axes. The largest and darkest circles on Figure 2.2 mark the strongest and most significant associations between characters and particular PCO axes. Broadly, there are more significant associations with the lower PCO axes, corroborating the results described above. This can be seen in two ways, either by noting that most of the dark circles are to the left of the plot, or by noting that both the height and darkness of the bars plotted beneath the x-axis increase to the left.

Regardless of the method used, the estimates all suggest that there is significant information contained in the PCO axes beyond the two or three dimensions that can be plotted practically. Such plots will provide a general indication of the arrangement of genera in morphospace rather than a comprehensive summary of the original data matrix. However, the observation that there is information in

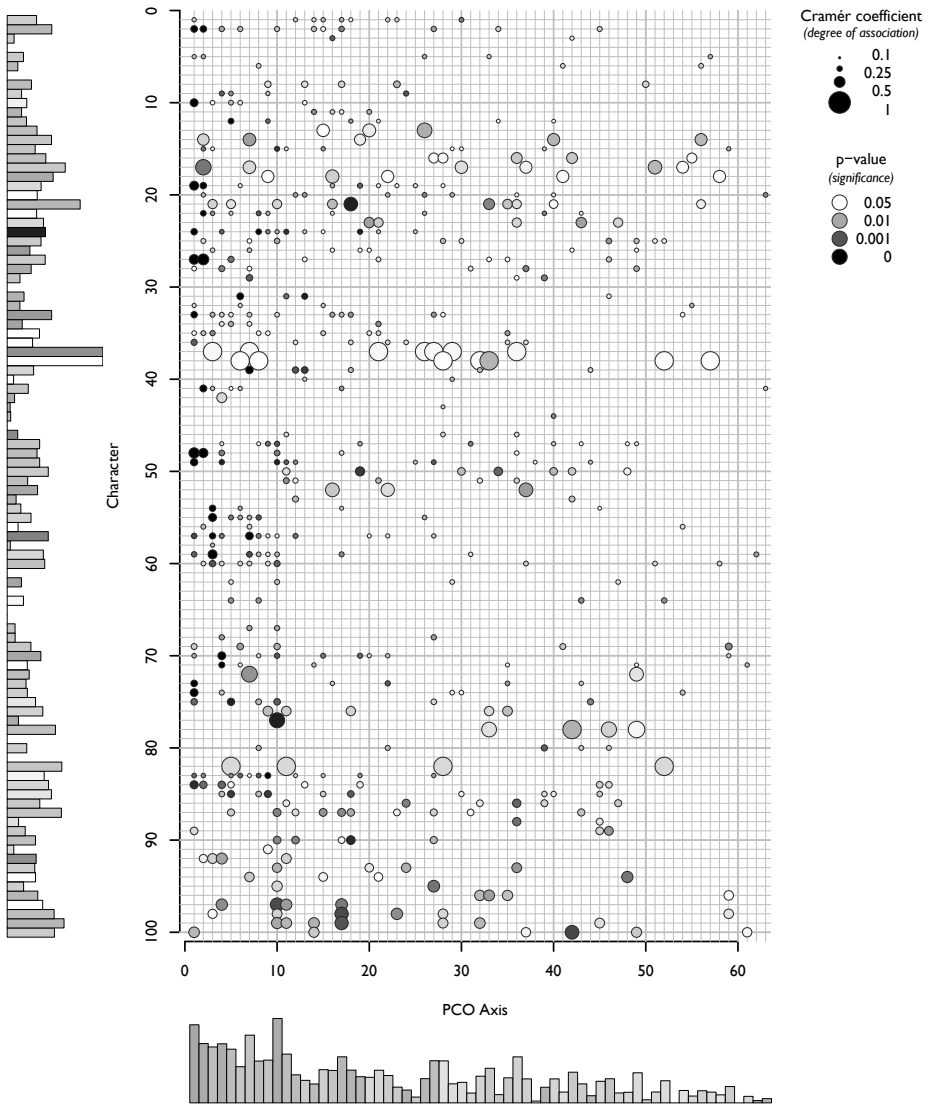


Figure 2.2: The degree of association between PCO axes (x-axis) and characters in the morphospace (y-axis). Circle diameter is proportional to the Cramér coefficient (from zero to one, zero suggesting the PCO score is independent of character state). Circle color indicates the associated p-value, darker meaning more significant. Comparisons with p-values > 0.05 were not plotted and were disregarded in marginal row and column sums.

higher PCO axes suggests there is important complexity in the original data set (as opposed to a handful of powerfully explanatory characters), and this suggests that a future effort to consider this information is warranted.

Interpretation of PCO Axes

Perhaps the most common criticism of ordinated or empirical morphospaces is that their axes are data-dependent (McGhee 1999; Wilson and Knoll 2010), but a related and more practical problem is that their axes are hard to interpret. Comparisons between theoretical and empirical morphospaces usually point to the distinction that the axes of the latter are unstable, with the dimensions changing upon addition or subtraction of more taxa, but what is more seldom mentioned is a related consequence of ordinating a high-dimensional space: the resulting axes represent a combination of many characters or parameters, making it difficult to understand what morphologies different parts of the ordinated space represent. In particular this restricts biologically meaningful interpretations of the morphospace, be they ecological, functional, or physiological (Wilson and Knoll 2010). In the following, we discuss three approaches to interpreting principal coordinate axes: labeling selected taxa with images, identifying which characters contribute most to the axes, and finally using the shape of the plot symbols themselves to represent states of morphological characters.

One widely used approach to understanding PCO axes is to use images highlighting some of the taxa (e.g. Swan and Saunders 1987, Fig. 1). This form of visualization (Fig. 2.3) suggests that pennate diatoms occupy the lower right quadrant, while bi- and multipolar centrics seem to be in the left half. While better than showing plot symbols alone, only some of the points can be labelled, leaving others unlabeled and thus with unclear morphologies—especially for large datasets. The images are also a complex composite of many morphological characters that can be difficult to deconvolve, because they cannot be located at the correct coordinates due to their size. Instead, representing the distribution of states of individual characters across the morphospace may give a clearer

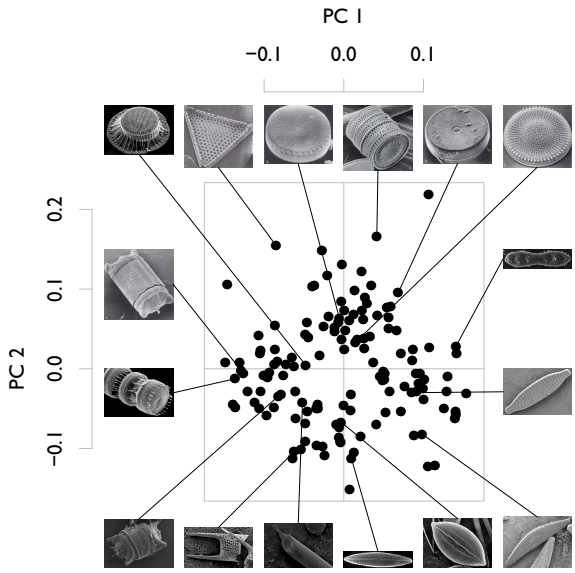


Figure 2.3: Morphospace plot of the first two PCO axes, annotated with images of selected taxa.

indication of what the axes mean.

We used the results of the character–PCO axis association summarized in Figure 2.2 above to identify which characters contribute most to the first two PCO axes used to visualize the morphospace (Fig. 2.3). Table 2.1 lists the characters with the strongest and most significant associations with PCO axes 1 and 2. Some of these characters are expected, particularly the shape of the structural pattern center of the primary silica ribs, because they are determinants both of overall morphology and of high-level taxonomy, and they thus reflect significant morphological variance. Other characters are more suprising, such as detailed features of the raphe or specialized processes, which apply to only a small subset of the genera in the analysis. A deeper statistical investigation would be needed to understand why characters we would expect, *a priori*, to be rather minor show such strong association with the first two PCO axes. However, it is plausible that characters with few states and many missing entries are simply more likely to fall into concordant patterns on the PCO axes by chance alone, in a

way that is not adequately corrected for in the calculation of p-values.

Table 2.1: The characters with the five highest Cramér coefficients and the five lowest associated p-values on the first two PCO axes.

Cramér coeff.	Axis	Char. #	Character description
0.84	PC 2	17	Central elevation shape
0.63	PC 2	14	Shape of apical elevation summit
0.59/0.53	PC 2/PC 1	27	Mantle shape in cross section
0.58	PC 1	100	Relative thickness of raphe sides
0.57/0.50	PC 1/PC 2	48	Shape of structural pattern center
0.49	PC 1	19	Angle between valve face and mantle
0.45	PC 2	92	Raphe extent
0.44	PC 1	84	Location of labiate process(es)

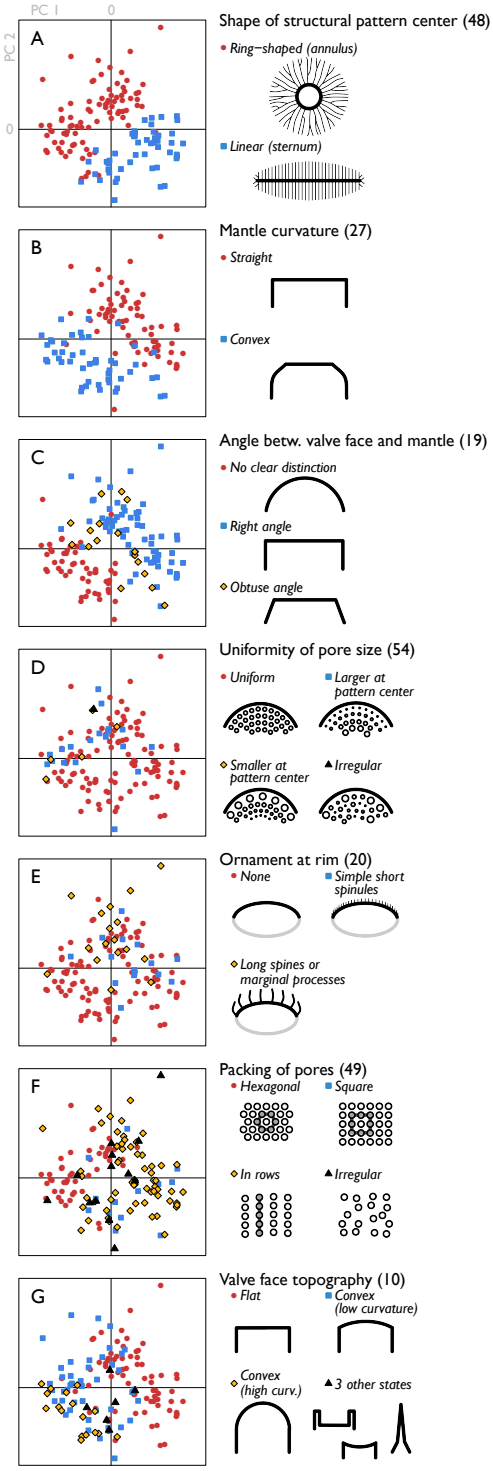
p-value	Axis	Char. #	Character description
$\leq$ 0.00001	PC 1/PC 2	19	Angle between valve face and mantle
$<$ 0.00001	PC 1	10	General topography of valve face
$<$ 0.00001	PC 1	49	Packing/coordination of pores
$<$ 0.00001	PC 1/PC 2	48	Shape of structural pattern center
$<$ 0.00001	PC 1	24	Depth of mantle
$<$ 0.00001	PC 2	27	Mantle shape in cross section
$<$ 0.00001	PC 2	41	Distinct central area
0.00003	PC 2	22	Marginal ridge at rim

We can get a somewhat clearer interpretation of what PCO axes 1 and 2 represent by plotting the different states of some of the characters most closely associated with those axes (Fig. 2.4). This exercise divides the plot area into clearly defined diagonal quadrants (Figs. 2.4A–C). Figure 2.4A confirms the suggestion from Figure 2.3 that centric forms lie in the upper left half, and pennate forms in the lower right half of the plot. Figures 2.4B–C, on the other hand, show an orthogonal division into forms with straight, clearly defined mantles in the upper right and forms with convex mantles without clear distinction from the valve face in the lower left.

The arrangement of character states in Figures 2.4D–G is less well defined, but still contributes meaning to the space defined by the two PCO axes shown. Diatoms with uniformly sized pores on the valve face occur all over the plot, while those with larger or smaller pores have positive PC 2 scores (Fig. 2.4D).

Figure 2.4 ***(following page)*****:** Morphospace plots of the first two PCO axes, with plot symbols denoting character states for seven of the characters (A-G, character numbers shown in parentheses, see Appendix B for detailed description) most associated with those axes (see Table 2.1 and Fig. 2.2).

Figure 2.4: (continued)



Similarly, diatoms with unornamented rims are found all over the plot, while those with short marginal spinules mostly have positive PC 1 scores and those with long marginal spines mostly have positive PC 2 scores (Fig. 2.4E). Most of the forms with valve face pores in hexagonal arrangement have negative PC 1 scores, while those in square arrangement or in rows tend to have positive PC 1 scores (Fig. 2.4F). Finally, the convexity of the valve face seems to decrease with increasing PC 1 score (Fig. 2.4G). In summary, Figure 2.4 reveals the following tendencies in the PCO space: (1) straight and clearly defined mantles toward the upper right versus indistinct and convex mantles toward the lower left of the plot, and (2) hexagonally-arranged pores and convex valve faces toward the left versus linearly-arranged pores and flatter valve faces toward the right of the plot.

We offer one further alternative visualization of the morphospace, using plot symbols generated from morphological character states to give a richer visualization of the morphospace than by using arbitrary symbols. Since the character states represent shape properties, it makes sense, at least in some cases, to use the shapes of the states as the plotting symbols themselves. We use the states of three characters describing the gross shape of the frustule to determine the form of the plot symbol (Fig. 2.5), showing a clear division between round and equant forms in the upper left and elongate forms, including raphe-bearing genera, in the lower right of the morphospace plot. This gives a more intuitive view of the notion, suggested by Figures 2.3 and 2.4, that centric diatoms occupy the upper left and pennates the lower right of the plot area. We also note that raphid diatoms, in this ordination, do not occupy an area distinct from the araphid pennates.

Armed with a visualization of the morphospace and a better understanding of its axes, we can begin to investigate the diatoms' evolutionary history. There are two major records of evolutionary history: the fossil record, and the record from genetic information. While we focus on the fossil record in this paper, we begin by briefly exploring the morphospace from the perspective of molecular data.

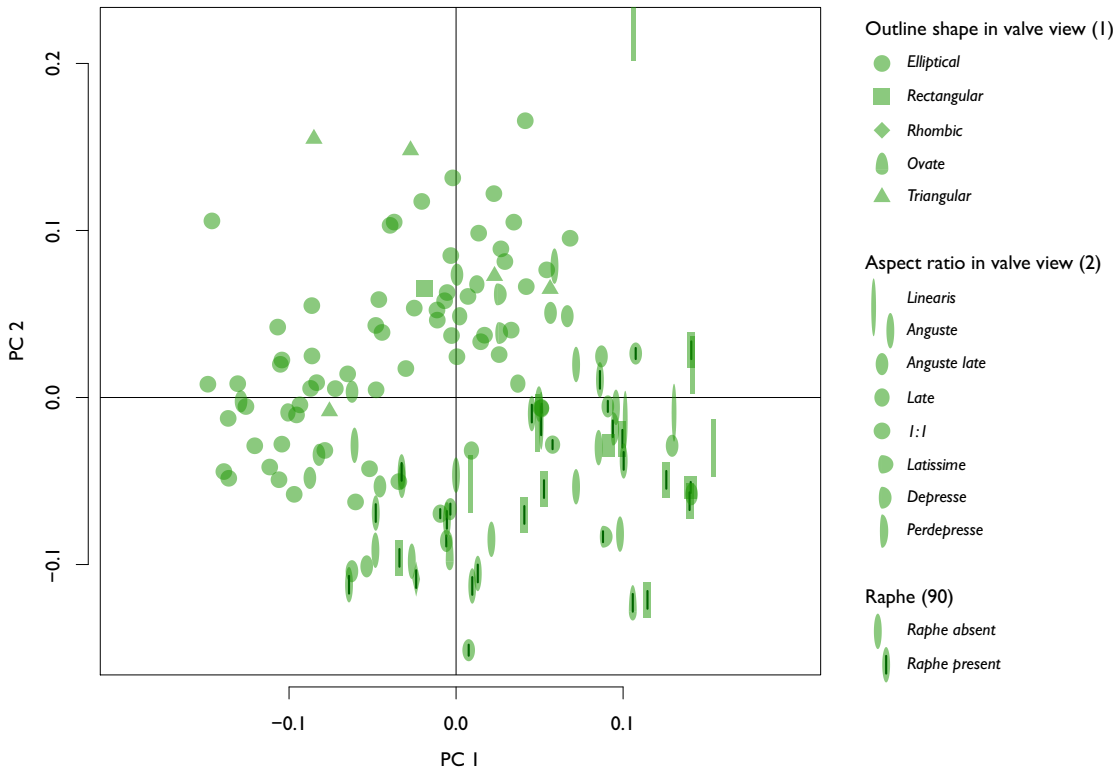


Figure 2.5: Morphospace plot of the first two PCO axes, with plot symbols generated from gross shape character states. The shape of the plot symbol—ellipse, rectangle, triangle, or oval—represents character 1 (the valve view outline shape category). The aspect ratio of the plot symbol represents character 2 (the aspect ratio of the diatom frustule in valve view). Character 90 (presence or absence of a raphe) is represented by a vertical line drawn within the plot symbol.

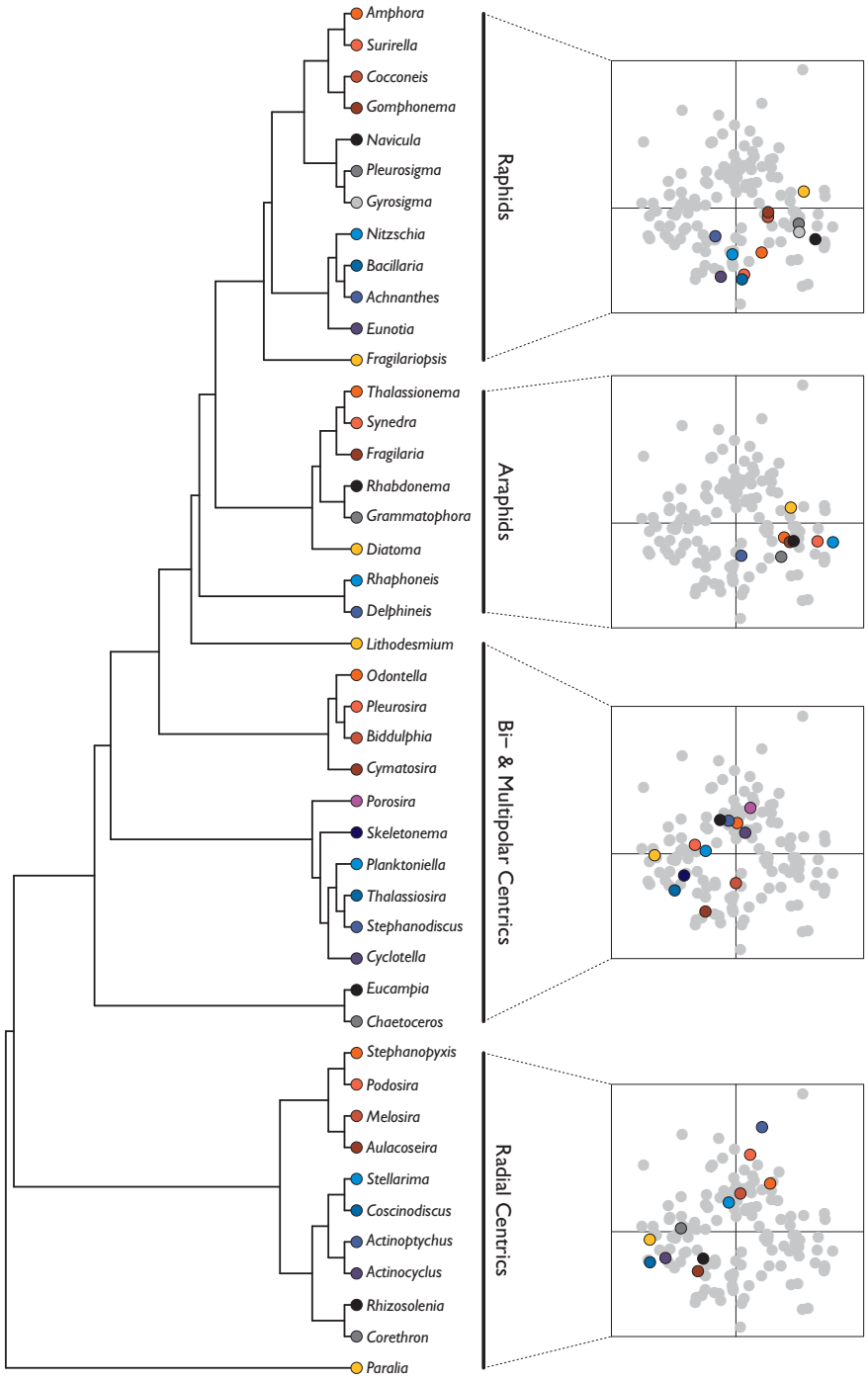
2.4.2 Morphospace and Molecular Data

What relationship between molecular phylogeny and morphospace would we expect to see if the Cenozoic Era was characterized by the occupation of significant new morphospace; in other words, if our expectation of Cenozoic stasis was false? If adding diversity were to add morphospace, we would see close relationships between the positions of genera on a tree and their position in morphospace, with derived clades occupying new, separate regions. More specifically, having identified the evolution of raphids and the Thalassiosirales as key Cenozoic events, we might expect these groups to occupy discrete regions of morphospace to the exclusion of other genera.

By comparing the distribution of genera on a phylogenetic tree with their distribution on a morphospace plot, however, we can see that only the coarsest phylogenetic division is reflected in morphospace (Fig. 2.6). The tree topology shown is a molecular phylogeny by Sorhannus (2007), based on a maximum-likelihood analysis of SSU rRNA sequences. Other molecular phylogenies give broadly similar results, though the detailed arrangement of genera varies (e.g. Kooistra et al. 2007). With the adjacent morphospace plots, the figure shows that pennates and centrics fall into different areas of morphospace (the lower right and upper left, respectively, as seen previously in Figs. 2.5 and 2.4), but groups at finer scales of phylogenetic resolution overlap. Within the pennates, for example, raphids and araphids fall in the same, overlapping region; radial and bi- and multipolar centrics also overlap. Clades within these groups do not occupy distinct regions to the exclusion of others; for example, the Thalassiosirales clade (*Porosira* through *Cyclotella* on the cladogram in Fig. 2.6, in various hues of blue/purple) do not fall in a distinct area within the bi- and multipolar centric group. This observation suggests that beyond the establishment of centrics and pennates, clades generally re-evolved the same gross morphologies and did not explore new and distinct areas of morphospace. It also suggests, in terms of gross morphology, that we cannot reject our hypothesis of stasis in morphospace occupancy after the radiation of pennate

Figure 2.6 ***(following page)*****:** Left, topology of a molecular phylogeny of diatoms (Sorhannus 2007) based on a maximum likelihood analysis of nuclear-encoded SSU rRNA sequences, trimmed to show only representative species from each of the 44 genera found both in the phylogeny and this study. The four plots on the right show where the genera in each of the four major groups fall in the morphospace (PCO axes 1 and 2, plot area as in Figs. 2.3, 2.4, & 2.5.) Within each of the four groups, genera are color-coded by proximity on the tree, e.g. in the top panel, the taxa colored red form a subclade within the raphids.

Figure 2.6: (continued)



diatoms, based on the interpretation of molecular data.

It is important to note that a number of other diatom phylogenies have been published based both on different taxa and on different tree building methods (e.g. Medlin and Kaczmarska 2004; Kooistra et al. 2007; for a review see Williams 2007). These phylogenies agree on the the largest-scale features of the diatom tree: centric diatoms form a paraphyletic group and pennates a clade, and similarly within pennates, the araphids form a paraphyletic group and the raphids a clade. Beyond this level, however, the phylogenies disagree on many details, and this variability suggests a level of uncertainty that complicates interpretation. The first-order observation that the major groups occupy overlapping regions of morphospace in this ordination, however, is not sensitive to these differences.

The lack of distinction in morphospace between araphid and raphid diatoms makes sense if we consider the function and ecological significance of the raphe. Because it allows for locomotion, the raphid diatoms are highly successful in terrestrial habitats, and the evolution of the raphe in diatoms has thus been compared to the evolution of flight in birds in its significance (Sims et al. 2006). However, because *Neptune* is mainly a deep-sea record of open-ocean plankton, the raphe may in fact be of limited significance in this environment, regardless of its overall importance to the group. Thus we might actually expect raphid pennates in the plankton to occupy the same functional and ecological niches as the araphid pennates, and—if form and function are related—that they thus occupy the same regions of morphospace.

The lack of correspondence between phylogeny and morphospace in Figure 2.6 might also be an artifact of the ordination of the morphospace. We have shown that much information is contained in higher PCO axes (Figs. 2.1 and 2.2), so we exercise caution in interpreting projected data directly. Fortunately, we can use the unordinated matrix of dissimilarities—i.e. the pairwise distances among genera in the full-dimensional space—to make a direct comparison with the phylogeny by calculating a comparable matrix of pairwise patristic distances (the sum of branch lengths, i.e. state changes along the branches, between two taxa) on the tree.

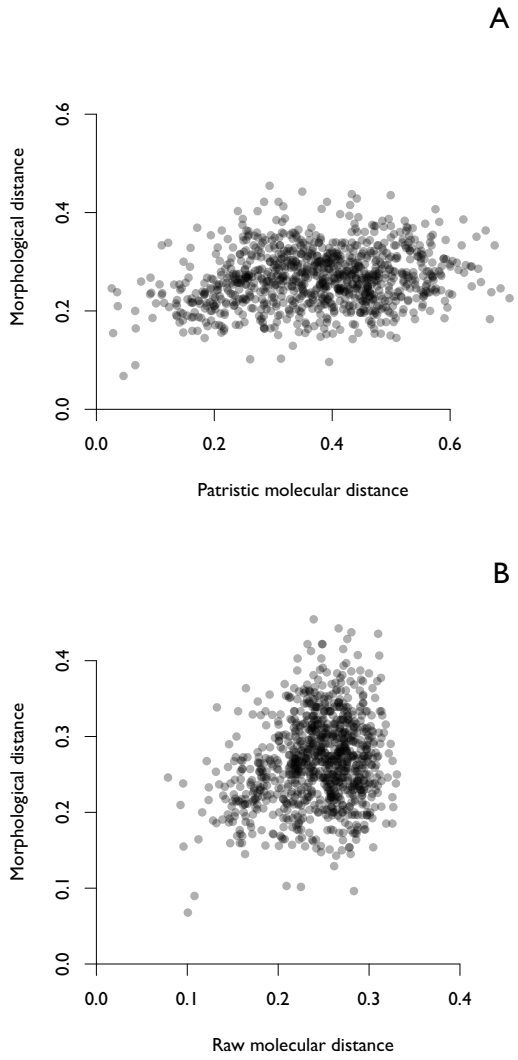


Figure 2.7: A, Pairwise morphological distances (character state mismatches divided by number of possible matches) plotted against patristic distance on the tree shown in Figure 2.6. B, Pairwise morphological distances plotted against pairwise molecular distance (identity between aligned sequences, calculated using the function `dist.alignment()` from the R package *seqinr*).

A direct comparison of morphological to patristic distance is shown in Figure 2.7A; it suggests very little correlation between the two. A simple linear regression of patristic distance on morphological distance has a squared correlation coefficient (R^2) value of 0.036, suggesting at most a very weak positive correlation. Interpreting the significance of this correlation is, however, complicated by the fact that we are comparing two distance matrices: a change in the position of any one point will cause the distances to all of the other points to change also, thus, the entries in each matrix are dependent on one another.

A statistical method designed specifically to test the correlation between two distance matrices is the Mantel test (Sokal and Rohlf 1981, p. 813). It is a type of permutation test in which one of the matrices is iteratively rearranged to generate a distribution of the correlation statistic to which the observed statistic can be compared. A Mantel test with 1,000,000 iterations gives a 2-sided p-value of 0.049, suggesting that there is a marginally significant relationship between patristic molecular distance and morphological distance at the 95% confidence level.

Rather than using patristic distances (Fig. 2.7A), we can compare morphological distances to molecular distance directly, using the distance between aligned molecular sequences (i.e., identity) in the absence of a phylogenetic hypothesis (Fig. 2.7B). Using molecular distance directly removes the subjective choices necessary in selecting tree-building methods. The R^2 in this case is only slightly higher, 0.057. The Mantel test for this comparison suggests that this relationship is also more significant, with a p-value of 0.024. If we accept these results, and if we assume that there is in fact an underlying positive relationship between morphology and molecular sequences, the somewhat surprising implication would be that phylogenetic tree-building actually masks that signal, weakening the correlation between the two sets of distances.

The qualitative sense provided by Figure 2.6 that the arrangement of taxa on the phylogenetic tree is not necessarily correlated with their arrangement in morphospace is thus confirmed quantitatively by a direct comparison of

morphological distance to molecular distance.

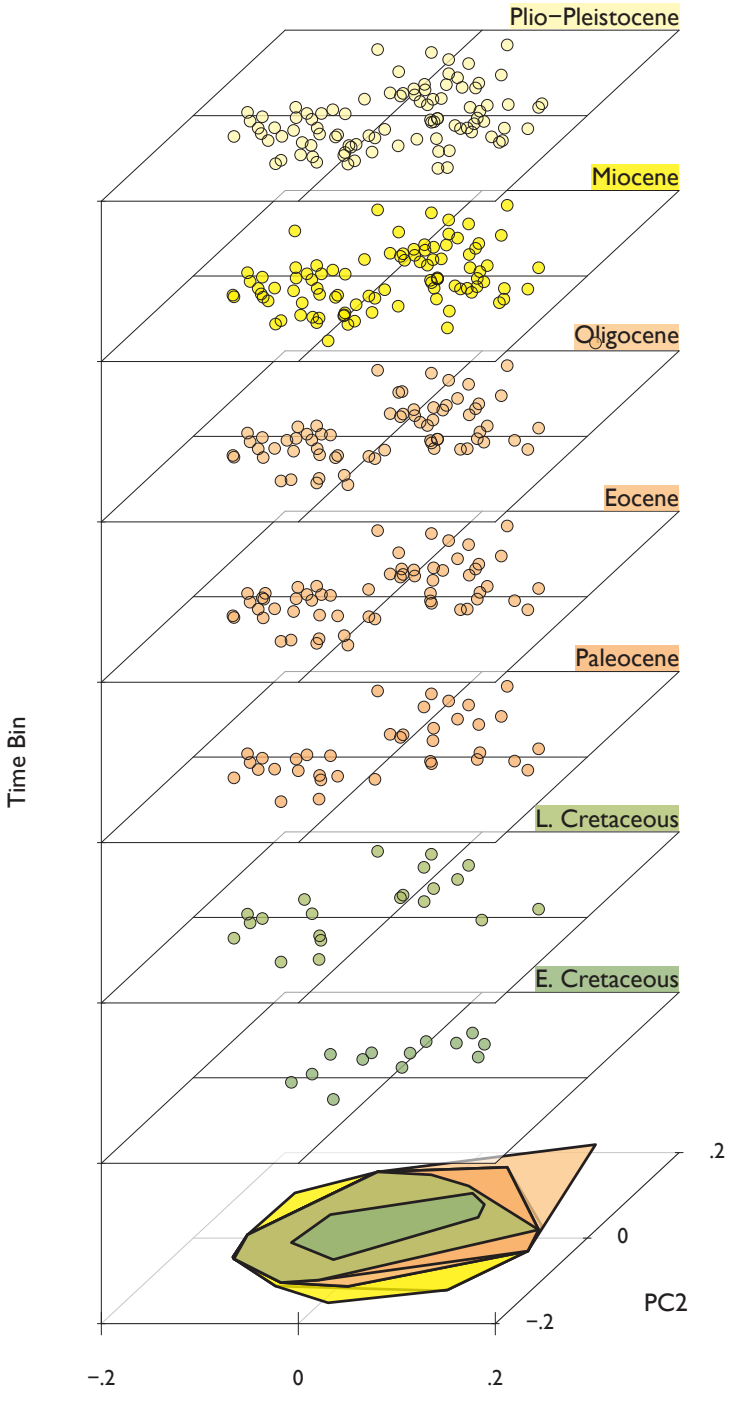
In summary, plotting phylogenetic relationships in morphospace suggests a weak relationship between morphology and descent. On the one hand, this is surprising, because diatom phylogenies predating the molecular era, and thus based on morphology, broadly agree with more recent molecular phylogenies. On the other hand, morphologically-based phylogenies rely on shared, derived features (synapomorphies) to signify inclusion in groups, while the data set underlying the morphospace consists of agnostically chosen, equally weighted (i.e. phenetic) characters. As such, we might not expect changes in the sequences coding for the ribosome to be correlated with frustule morphology, on which those sequences presumably have little direct bearing. Expected or not, the results of comparing phylogeny and morphospace suggest that different groups of diatoms, and subgroups within those groups, successively recolonized already-occupied regions of morphospace. Since the four major groups were already present by the earliest Cenozoic Era, the full extent of occupied morphospace should have been achieved early, and show little subsequent change. These results support the hypothesis that, in terms of disparity or morphological variety, the Cenozoic Era was broadly characterized by stasis.

2.4.3 Morphospace Through Time

We now explore occupancy of the morphospace through the other record of diatom evolutionary history, the fossil record. When viewed as Cenozoic epochs in PCO axes 1 and 2, the occupied morphospace area seems relatively constant through time, to a first approximation (colored polygons at the bottom of Fig. 2.8). The area occupied appears to expand slightly to the lower right and upper left by the Miocene. The Oligocene area is expanded to the extreme upper right, but this is due to a single taxon with an unusual morphology (see point between "O" and "l" of "Oligocene"). In addition to the slight expansion of morphospace area, sparsely occupied areas appear to become "filled in" and more densely occupied through time.

Figure 2.8 *(following page)*: Morphospace, as represented by the first two PCO axes, resolved through time using range-through taxon counting of *Neptune* occurrences. The colored polygons at the bottom of the plot are convex hulls enclosing the taxa present at each time bin, labeled in the corresponding colors.

Figure 2.8: (continued)



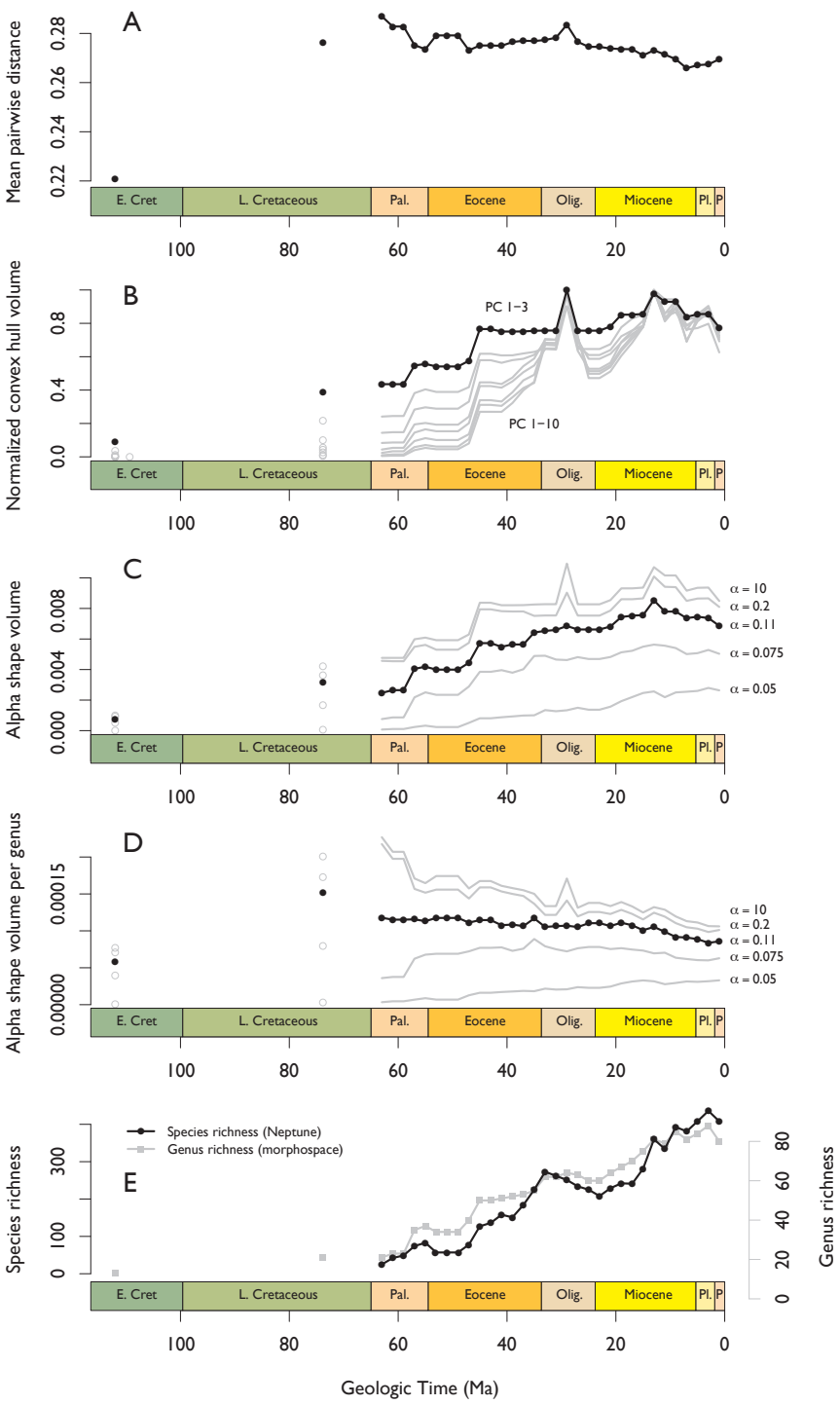
The Cretaceous time bins, particularly the Early Cretaceous, appear to occupy a much smaller area of morphospace. However, rigorously interpreting the Cretaceous results is challenging because so much less data was included than in the *Neptune* database. Specifically, the Early and Late Cretaceous time bins contain taxon lists from one and three ODP holes respectively, while the Paleocene alone contains lists from 61 samples from six ODP holes. Furthermore, several morphologically divergent taxa did not meet the applied culling threshold, due to incomplete descriptions, and were thus excluded from the analysis. The Cretaceous samples may thus show less morphological variety than was actually present, though it was probably still lower than the Cenozoic samples, particularly the Early Cretaceous.

There are numerous ways of quantifying disparity, or what has been called the "within-group variance of form" (Erwin 2007), that go beyond the qualitative description of morphospace occupancy provided by plots like Figure 2.8. They including counts of higher taxa, the sum of univariate variances, total range, the number of unique pairwise character combinations, participation ratio, various measures of PCO volume, and mean pairwise distance (for details, see Thomas and Reif 1993; Foote 1995a; Ciampaglio et al. 2001; Erwin 2007). Some of these metrics, as explained below, may describe different aspects of morphospace occupation, two major aspects of which are how far taxa are from each other, on average, and what volume of the space is occupied. Next, we present metrics for those two aspects, using mean pairwise distance to describe the former, and two measures of occupied PCO volume (convex hull and alpha shape volume) to describe the latter.

Mean pairwise distance is a commonly used metric for disparity (for example, by Foote 1995a; Lupia 1999; Boyce and Knoll 2002), having the advantage that it can be calculated from the morphological data directly without requiring ordination. Another advantage of this metric is that it has been shown to be relatively insensitive to sampling bias (Foote 1995a; Ciampaglio et al. 2001; Deline 2009). Mean pairwise distance suggests that disparity changed little over the course of the Cenozoic Era, showing a slight decline (Fig. 2.9A). These

Figure 2.9 *(following page)*: Metrics of morphological disparity (A-D) and diversity (E) through time, using *Neptune* occurrences under range-through taxon counting. A, Mean pairwise dissimilarity between genera, as character state mismatches divided by number of possible matches. B, Convex hull (hyper-)volume containing genera, normalized to largest value; black line is volume calculated over the first three PCO axes, grey lines are volume over the first four, five, etc. up to ten PCO axes. C, Alpha shape volume containing genera; black line is volume for α-value chosen by inspection to best capture occupied volume across time bins, grey lines are other α-values. $\alpha = 10$ recovers the convex hull solution. D, Alpha shape volume (as in C) divided by number of genera. E, Species-level diversity from *Neptune* database (includes genera left out of morphospace analysis) in black; genus-level diversity in morphospace analysis in grey.

Figure 2.9: (continued)



results show that pairs of genera are, on average, about 70–75% similar in applicable characters, with an apparent peak in the Oligocene and declining gradually over the course of the Cenozoic Era. A disadvantage of this method is that is says nothing about the total extent or shape of the occupied morphospace.

Calculating convex hull volume, another disparity metric, is a way of quantifying the amount of space occupied by a set of points (Foote 1999). A convex hull is a shape enclosing a set of points using the smallest possible number of those points (in two dimensions, it is the equivalent of spanning a rubber band around a set of pegs). The volume (or hypervolume) of this shape for each time bin was calculated for increasing numbers of PCO axes, up to 10 (beyond which computational limits are reached). In order to be comparable, the results are presented standardized to the largest value in the time series.

The convex hull volumes calculated are shown in 2.9B. The plot shows an increase in volume with time, regardless of the number of dimensions used to calculate it. There is a decline in volume over the most recent 5 Myr or so; however, this may be related to the well-known edge effect of the range-through taxon counting method (Raup 1972; Alroy 2010a). The largest volume is reached in the Oligocene, showing a particularly pronounced spike in the 29 Ma time bin. However, by examining the Oligocene time slice plotted in Figure 2.8, it is clear that this spike is due to a single outlier taxon present only at that time. This illustrates a shortcoming of the convex hull method: due to outliers or widely separated clusters of points, it can include substantial areas of unoccupied space.

Alpha shapes are a generalization of convex hulls that, when appropriate values of α are chosen, address the empty-space problem using the convex hull method of quantifying morphospace occupation. Alpha shapes (Edelsbrunner and Mücke 1992) allow unoccupied space to be removed from the convex hull, akin to "scooping out" space with an ice-cream scoop of a given radius, α; as the value of α increases, the alpha shape converges on the convex hull. The method was first applied to morphospaces by Low (2006). We used the *alphashape3d* package in *R* (Lafarge and Pateiro-Lopez 2012); it is limited to calculating volume in three dimensions. From this exercise we find that the morphospace occupation

shows the same pattern of secular increase in volume as the convex hull volume, but without the exaggerated peaks (Fig. 2.9C). Alpha shape volume roughly doubles over the Cenozoic Era.

These different metrics of disparity—mean pairwise distance and the volume of morphospace occupied—give very different results because they measure different aspects of disparity. Mean pairwise distance declines slightly over the Cenozoic Era, in stark contrast to occupied volume (as calculated either by convex hulls or by alpha shapes), which increases substantially over time. These divergent results can be understood as measuring two different aspects of morphospace occupation. The volume increases as the extent of morphospace occupied increases. If the number of genera were to stay constant, we would also expect a concomitant increase in average pairwise distance. However, the number of genera occupying this space also increases through time, leaving genera packed more tightly into morphospace and thus reducing the average distance between them. Disparity can thus both increase substantially and decrease slightly over the Cenozoic Era—the former in the sense of the range of morphological variety, and the latter in the sense of the average morphological distinctness of taxa.

Another way to quantify the "packing" of morphospace suggested by the decline in mean pairwise distance is to calculate the volume of morphospace occupied per taxon. This result is shown in Figure 2.9D and it shows a similar trend to the mean pairwise distance results in Figure 2.9A: the amount of PCO volume per genus is slightly decreasing, pointing again to an increase in the number of taxa filling morphospace outpacing the growth of the volume occupied.

Because the higher PCO axes contain substantial information (Figs. 2.1 and 2.2), we noted that results based only on a few ordinated axes should be interpreted with caution. In order to examine morphospace occupancy in a more direct way, we counted the number of realized character states through time, considering the raw morphological data without ordination. This metric is similar to the number of realized unique pairwise character combinations of Thomas and Reif (1993) and Foote (1995a), and is of a lower dimensionality

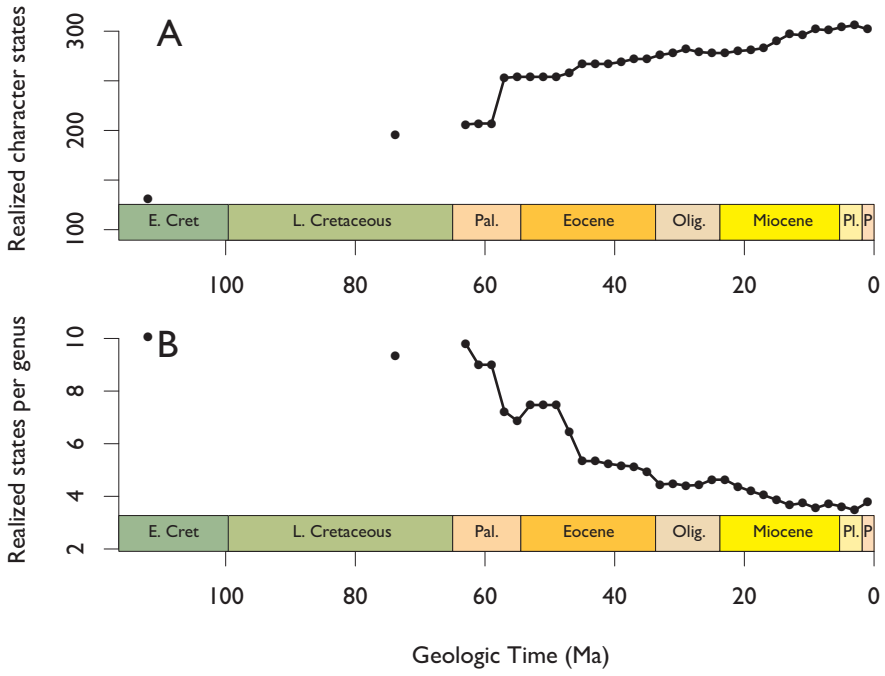


Figure 2.10: Number of morphological character states observed through time. A, Number of realized states. B, Number of states (as in A) divided by the number of genera (as in Fig. 2.9E). The total number of states in the (culled) morphological data matrix used in the analysis is 317.

than the extremely sparsely populated full morphospace. However, it considers only whether a character state is realized, independent of other characters. In morphospace studies with relatively few characters, the former approach is preferable because the 1-dimensional space of character states can quickly become saturated (i.e., all character states are, more or less, always realized, but occur in different combinations in different taxa). In the present study, however, the space of character states only approaches saturation at the very end of the time series (Fig. 2.10A), and it therefore has sufficient sensitivity to render pairwise comparisons unnecessary.

Figure 2.10A shows that the number of realized character states increases through time, agreeing with the results from the PCO volume metrics, and confirming that the range of occupied morphospace expands through the Cenozoic Era. However, as we have seen in Figure 2.9E, the number of genera also increases over that time. Figure 2.10B shows the number of realized character states divided by the number of genera, a metric that decreases by more than a half over the Cenozoic Era. We interpret this to mean that as new taxa evolved in the Cenozoic Era, they increasingly showed new combinations of existing character states over newly evolved states, even as new states continued to evolve. Another way to understand this result is to consider it as a decrease in the amount of morphospace unique to each taxon. In both ways, this result mirrors the slight decline in the mean pairwise distance result shown in Figure 2.9A. The concordance of these two sets of results from ordinated and unordinated morphospace data (PCO volume occupied agreeing with number of realized states, and mean pairwise distance agreeing with per-genus realized states) lends confidence to our interpretations from ordinated data.

A final aspect to consider concerns the increase in sampling intensity over the Cenozoic Era, which casts doubt on the reliability of the observed increases in morphospace occupation over that time period (Fig. 2.11). The roughly exponential increase in sampling raises the question whether the observed increases in morphospace occupation (as seen in Figs. 2.9B-C and Fig. 2.10A) are real or result from sampling biases. The importance of secular variation in

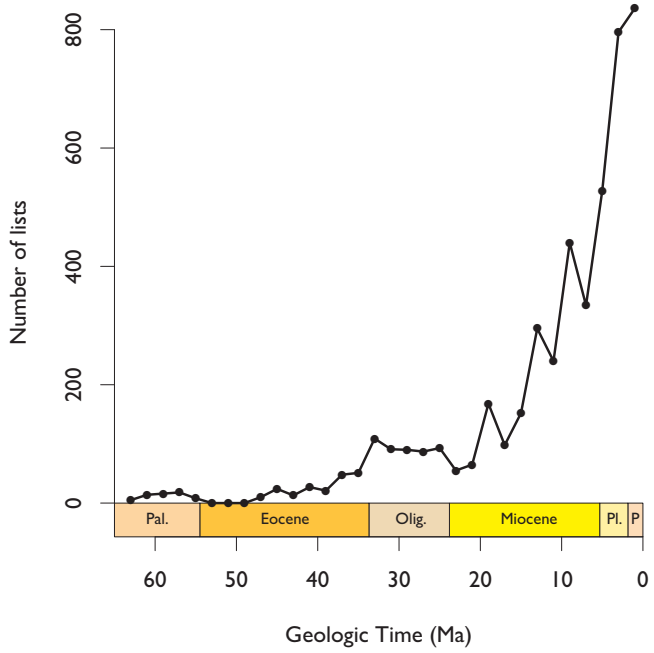


Figure 2.11: Number of taxonomic lists (i.e. described samples) per time bin in the *Neptune* database. Sampling increases approximately exponentially with time.

sampling intensity is well established in studies of taxonomic diversity through time (e.g. Alroy et al. 2001), where sampling biases have been shown to (1) greatly attenuate patterns of diversity increase, and (2) shift the timing of peaks, or even reverse patterns (reviewed by Alroy 2010a). The *Neptune* record has been widely cited as the canonical compilation for diatom diversity, but its uneven sampling has been identified and attempts at correcting for it have been made by applying sampling standardization methods (Rabosky and Sorhannus 2009). We tend to think morphospaces and studies of morphological disparity constitute a window to evolutionary history that is independent of taxonomic diversity, and this may, in part, explain why sampling biases have often not been considered (see, however, Foote 1995a; Ciampaglio et al. 2001; Shen et al. 2008; Deline 2009). While two data sets of taxonomic diversity and morphological disparity do indeed offer different information, both are subject to the same underlying sampling biases. These biases are considered in Chapter 3.

2.5 Conclusions

Diatom morphospace can be visually depicted using plot symbols whose shape reflects morphological characters of the taxa they represent. This alternative to plotting generic symbols, like dots or crosses, or labeling selected points with images, goes some way towards correcting the shortcoming of many empirical morphospaces that lack clear identification of what their axes mean.

Plotting phylogenetic relationships onto diatom morphospace suggests very little relationship between morphology and descent; this implies that the same regions of morphospace were iteratively colonized by different clades. Thalassiosirales and raphid pennates—clades that evolved in the Cenozoic Era—do not appear to occupy regions of morphospace distinct from the clades within which they arose. From the phylogenetic perspective, then, most of the extent of morphospace seems to have been occupied early, suggesting that the Cenozoic Era was broadly characterized by stasis in terms of disparity, or morphological variety.

We examined changes in Cenozoic diatom morphospace occupation through time using the *Neptune* database, based on the marine fossil record, and calculated disparity in each time slice. Two sets of disparity metrics show different secular trends, which we argue is a consequence of the fact that they measure different aspects of disparity.

The "packing" of morphospace, or how much morphospace on average separates taxa, can be measured using mean pairwise distance, the per-genus alpha shape volume, or the per-genus number of realized character states. The last shows a decreasing trend, while the first two show only a slight decline through the Cenozoic Era, varying somewhat with the choice of the α parameter.

The volume of morphospace occupied, delimited by convex hulls or alpha shapes (the latter are less distorted by outliers), and the number of realized character states are proxies for the total extent or amount of morphospace occupied. These volume metrics both show an increase through the Cenozoic Era.

Taken together, these metrics as showing an increase in the total extent of occupied morphospace, with an associated increase in the number of taxa keeping pace with the rate of space expansion, which leads to stationary or even increasing "packing" of taxa through Cenozoic time.

A number of lines of evidence, then, point to Cenozoic morphological stasis: mean pairwise distance, alpha volume per genus, and the phylogenetic view of morphospace. In contrast, measures of the total extent of occupied morphospace, when viewed independently, suggest an increase through time. We suspect, however, that the latter are affected by sampling bias, as suggested by a corresponding increase in the number of taxa in the morphospace analysis and the number of taxonomic lists in the *Neptune* database.

Since mean pairwise distance has been shown to be relatively insensitive to sampling bias, we believe that our results point toward Cenozoic stasis in overall morphology. This conclusion could be further substantiated by applying sampling-standardization methods, such as those developed for studies of taxonomic diversity, to diatom morphospace.

3

Sampling standardization and sensitivity tests of planktonic marine diatom morphospace

Abstract

The occupation of morphospace through time provides a morphological view of diversification distinct from the more familiar taxonomic tabulations. However, this view is subject to the same geological biases long recognized in studies of taxonomic diversification, where techniques for correcting secular bias in sampling have become standard practice. In this study, we apply techniques of sampling standardization to a morphospace investigation of diatoms in order to test whether the observed stratigraphic trends in morphospace occupation are artifacts of secular trends in

sampling. When sampling bias is corrected by randomized subsampling, all disparity metrics show stationary patterns, or at most directional changes of small magnitude. We find that disparity metrics describing the average dispersion of taxa in morphospace are less subject to sampling bias than those describing the total extent of morphospace occupied. We also investigate a measure of disparity that is insensitive to sampling intensity, introducing a geographic component of morphological disparity. By analogy to α and β components of taxonomic diversity, we suggest the notions of α and β disparity. We find that $\bar{\alpha}$ disparity remains roughly constant through time. As a by-product of applying subsampling methods to diatom morphospace, we present the first taxonomic diversity curve of diatoms under shareholder quorum subsampling (SQS) and find results similar to those of previously published subsampling methods: a roughly twofold rise over the Cenozoic, with peak diversity around the Eocene/Oligocene boundary. In testing for methodological bias from choices in ordination method and data culling during morphospace construction, we find our results are relatively insensitive to both factors. Our results show that the Cenozoic occupation of planktonic diatom morphospace is largely characterized by stasis. More broadly, our results make clear that a complete view of morphological disparity must consider sampling biases, which can be addressed with well-established, quantitative methods in morphospaces populated using occurrence-level data.

3.1 Introduction

The study of the fossil record makes valuable contributions to our understanding of evolution, not least through documenting the diversification history of clades. By analyzing the occupation of morphospace through time, we can compare a morphological perspective on diversification (through metrics of disparity) to the more familiar taxonomic view of diversification (through counts of species richness). Many of the groups in which this comparison has been made show "asymmetric diversification" where peak morphological disparity is reached early and then remains more or less stationary while taxonomic diversification continues (Gould 1989; Foote 1997; Erwin 2007). In marine planktonic

diatoms, an ecologically important group of primary producers with siliceous cell walls called frustules, the history of taxonomic diversification has received more attention than morphological diversification. Their taxonomic diversification history has conventionally been read as a sharp Cenozoic rise to current levels of diversity, relatively late in an evolutionary history stretching back to at least the Early Cretaceous Period.

In Chapter 2, we addressed the Cenozoic history of diatom morphological disparity through a study of their fossil record. We showed that as taxonomic richness increased, the range of morphospace occupied increased also, while the distance between taxa in morphospace was characterized by stasis or even gradual decline. We stopped short of making strong biological inferences from these observations, however, because of the possibility that these results are subject to biases, particularly from temporally uneven sampling. Such biases have long been recognized in studies of taxonomic diversification, in which the impact of different methodological choices has been investigated (e.g., different taxon-counting methods, Bambach 1999), and techniques for correcting for secular bias in sampling, like rarefaction, by-list subsampling, or shareholder quorum subsampling have become standard practice (Miller and Foote 1996; Alroy et al. 2001; Alroy 2010b). The impact of bias (particularly from sampling), however, is not as often considered in studies of morphospace, even though trends in morphological disparity are commonly compared to taxonomic diversity. This might be explained by the widespread recognition that morphological data represent a different window into evolutionary history than taxonomic data, perhaps distracting from the fact that both are derived from the fossil record, and are thus both subject to the well-known geological biases that have been the subject of research since the origins of the discipline (e.g. Darwin 1859; Newell 1959; Raup 1972).

Nonetheless, Foote (1992) did recognize the importance of sample size in assessing morphological disparity and applied rarefaction to metrics of morphospace occupation in trilobites, blastoids, and ammonoids. However, Foote's definition of a "sample"—the unit being rarefied or subsampled to a

common threshold—in that study is quite different from the definition in current studies seeking to correct for sampling bias in time series of taxonomic richness. In the study by Foote (1992), each *taxon* in the morphospace in a given time bin is considered a sample, while in diversity subsampling studies, each *occurrence* of a taxon (or assemblage of taxa) is considered a sample. Thus, rarefied time bins in Foote's morphospaces contain the same number of taxa, while in studies of taxonomic diversity, rarefied time bins contain the same number of occurrences (e.g. Miller and Foote 1996).

In a more recent morphospace study of the Ediacara biota, Shen et al. (2008) approached the problem of sampling bias by calculating a metric of morphospace occupation under rarefaction using the latter definition, treating taxon occurrences as samples. However, these authors report the results of rarefaction for just one of three time bins, and do not attempt to correct comprehensively for sampling differences. We are not aware of any morphospace study to date in which sampling differences have been corrected by sampling standardization as has become common practice for studies of taxonomic diversity.

The need to correct for uneven sampling of the fossil record in studies of morphological diversification was recently highlighted in a study of pterosaur disparity (Butler et al. 2012). The authors demonstrate significant correlations between proxies of geological sampling and metrics of morphospace occupancy and conclude that disparity metrics based on the range of occupied morphospace, in particular, are strongly affected by uneven sampling of the fossil record. Although they do apply rarefaction to standardize disparity metrics, it is subsampling of the sort performed by Foote (1992), to a standard number of taxa. Although occurrence-level data are available, due to the nature of the pterosaur fossil record—in which almost every occurrence is a singleton (i.e. the only occurrence of that taxon)—no meaningful sampling standardization on the level of occurrences (in the sense of diversity studies like Alroy et al. 2008) is possible.

Temporally uneven sampling was identified as a possible confounding factor in interpreting the results of the diatom morphospace study in Chapter 2. Because

the diatom fossil record can yield many thousands of individuals in a spoonful of sediment, and since the *Neptune* database of microfossil occurrences (Lazarus 1994; Spencer-Cervato 1999) captures much of this information, we can directly address sampling biases here.

In this study, we extend the techniques of sampling standardization developed for studies of taxonomic diversity history by applying them to a morphospace of diatoms in order to test whether the results presented in the Chapter 2 are artifacts of secular trends in sampling. We use various subsampling methods including the recently published shareholder quorum (SQS) method; in the process, we also report the first application of SQS to the diatom record of taxonomic diversification. We further test for sampling bias by examining disparity metrics that ought to be insensitive to sampling differences. We also test for methodological bias in constructing the morphospace, from choices in ordination method and the choice of thresholds for data culling based on missing information. Finally, we look for biological signals in the data by examining the distribution of sets of characters expected to change under suggested drivers of macroevolutionary change over the Cenozoic Era.

3.2 Materials and Methods

3.2.1 Morphospace Construction and Disparity Metrics

We constructed an empirical morphospace (McGhee 1999) by coding the states of 123 discrete binary or unordered multistate morphological characters for 152 diatom genera. The chosen genera represent all valid genera in the *Neptune* database, less those identified as resting stages. This choice of taxa made it possible to use the *Neptune* database to populate the morphospace through time and apply sampling-standardization methods at the level of occurrences. We use principal coordinates analysis (PCO) to transform the data to continuous form, and binned occurrences into 2-Myr time intervals to calculate four disparity metrics describing the occupancy of this morphospace: the volume of the convex

hull encompassing the taxa present, the volume of an alpha shape encompassing the taxa present, the alpha shape volume divided by the number of taxa, and the mean pairwise distance (measured as the number of character state mismatches divided by the number of possible matches). The first two measure the total amount of morphospace occupied, while the last two measure how close taxa are to one another in morphospace.

All analyses were carried out in the statistical programming language *R* (R Development Core Team 2011); the software written to carry out the analyses and produce the figures shown is provided in Appendix F. A detailed description of the method of morphospace construction, including the choice of morphological characters and the calculation of disparity metrics, is given in Chapter 2.

3.2.2 Morphospace Subsampling

We carried out sampling standardization using four different subsampling methods: "classical" rarefaction (CR, Miller and Foote 1996), by-list unweighted (UW, Alroy et al. 2001), by-list weighted by occurrences (OW, Alroy 1996), and shareholder quorum subsampling (Alroy 2010b). These methods are reviewed in detail by Alroy (2010a) and are only briefly outlined here.

In each of these methods, occurrences are drawn from the full dataset until a given quota is reached. Morphospace metrics are calculated on this subsample and the process is repeated many times; the mean and confidence intervals of these iterations are reported. In CR, occurrences are drawn individually until a quota of a number of occurrences is reached. In UW, occurrences are drawn by taxonomic list (a list of taxa reported from one slide at one depth in one borehole) until a quota of a given number of lists is drawn. In OW, occurrences are also drawn by-list, but the quota is a given number of occurrences. These subsampling methods are the same as those carried out by Rabosky and Sorhannus (2009), although we do not apply O2W subsampling (in which the quota is a sum of squared occurrences) due to the strong biases in that method

when beta diversity is non-negligible, as demonstrated by Bush et al. (2004). Also, since we require a list of taxon names present in each subsample—rather than just the number of taxa—in order to calculate metrics of morphospace occupancy, we do not apply a sampling probability correction (the "three-timer" correction of Alroy et al. 2008).

These methods of sampling standardization seek to achieve uniform sampling through time, but Alroy (2010a,b) has argued that uniform sampling is not necessarily fair sampling. He suggested that fair sampling should sample the same proportion of total diversity in each interval—meaning that more diverse intervals will require more sampling than less diverse intervals to recover an accurate diversity curve. He proposed a new sampling standardization method, shareholder quorum subsampling (SQS), which hinges upon estimating the proportion of total diversity represented by a sample. This is achieved using Good's *u* (Good 1953), a metric from ecology that uses the prevalence of singletons in a sample as an indication of coverage. Alroy (2010a,b) modified this for use in SQS by substituting taxa occurring in a single publication in place of singletons. The *Neptune* database, however, does not include direct information about source publications, and in any case, the source publications rarely contain singleton occurrences because of the way micropaleontological data are collected (they report occurrences of a set of taxa over a stratigraphic range). We thus apply a further modification to this estimate, substituting for single-publication taxa the number of taxa occurring in only one DSDP/ODP borehole. We also neither apply the largest collection correction nor do we discard the most abundant taxon in each sample because we do not consider the related biases to apply to the *Neptune* data. Finally, the current version of our software does not implement the "throwback" refinement of Alroy (2010a,b), meaning that each subsample will have a quorum level slightly exceeding the target.

Because our morphospace is constructed at the genus level with some taxa excluded (see Chapter 2), we report both the genus richness recovered by the morphospace subsampling exercises as well as the species richness obtained in separate subsampling of the complete *Neptune* data.

3.3 Analysis

3.3.1 Distribution of Occurrences in Morphospace

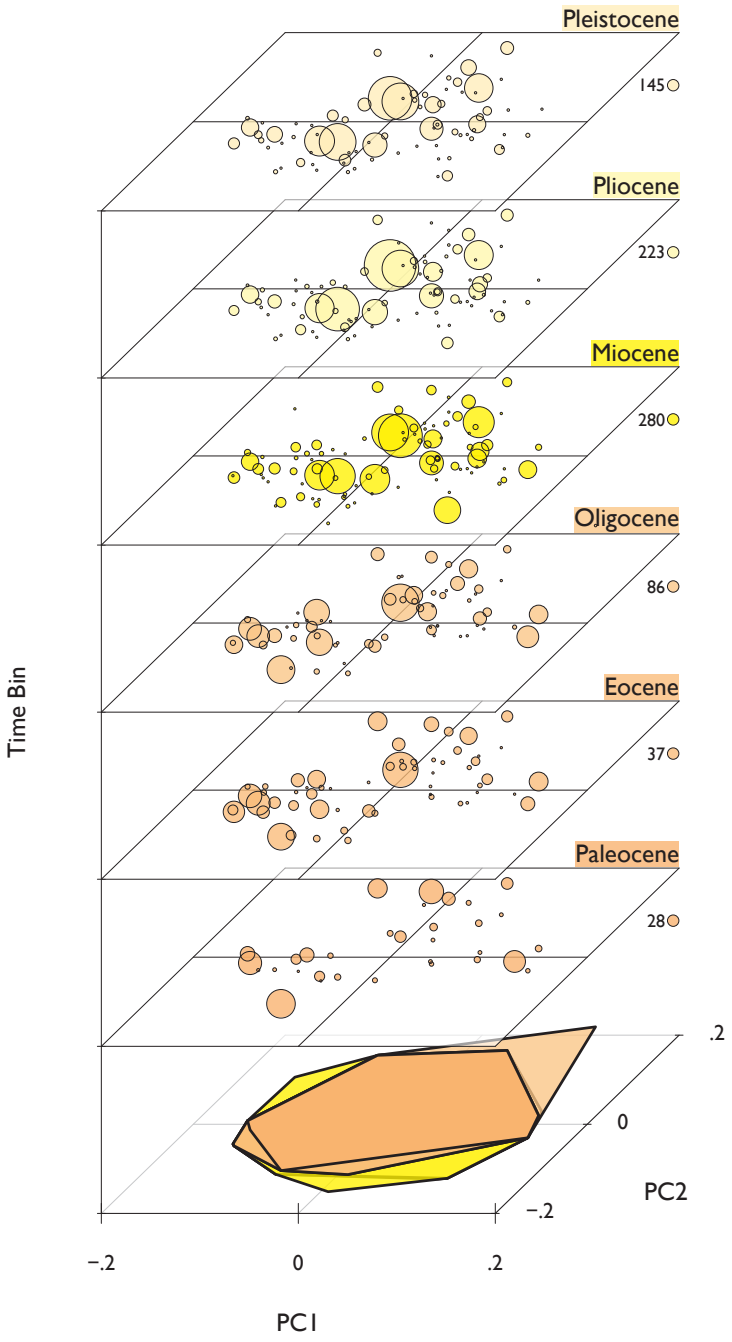
Visualizations conventionally display the occupation of morphospace in a binary fashion: a point in morphospace is either occupied by a taxon—shown by a point plotted in the chosen ordination at the location representing the taxon's morphology—or it is not (e.g. Fig. 2.3). When an occurrence-level database is used to populate the morphospace, however, an additional dimension of information can be shown by representing the number of occurrences of each taxon by the size of each plotted point.

Plotting Cenozoic diatom morphospace occupation in this way shows that some areas of morphospace are more sparsely occupied than others in terms of fossil occurrences (Fig. 3.1). In Chapter 2, we pointed out the possibility that the Cenozoic rise in the number of *Neptune* occurrences might bias our metrics of disparity. Figure 3.1 gives a more nuanced view of the need to consider sampling differences, making it clear that some regions of morphospace are occupied by few occurrences. Had the younger intervals been sampled less, at a level comparable to the Paleocene, those regions may have been seen as unoccupied.

The same observation could have been made by simply comparing rank-ordered abundance distributions for different *Neptune* time bins. But Figure 3.1 suggests something further: that these occurrences may not to be randomly distributed in morphospace, at least as viewed through the first two PCO axes. In the Paleogene time bins, taxa defining the edges of occupied morphospace appear to have relatively many occurrences. In contrast, in the Neogene (and particularly the Plio-Pleistocene), the edges of morphospace are largely occupied by taxa with few occurrences. This observation calls into question the interpretation of disparity metrics based on the range or volume of morphospace occupied (Figs. 2.9B and C). It suggests the possibility that, under sampling comparable to older time bins, the younger time bins may not have shown the observed increase in the total extent of occupied morphospace.

Figure 3.1 ***(following page)*****:** Morphospace plot of the first two PCO axes through time, with the size of each plot point representing the number of occurrences of that taxon in the *Neptune* database. Plot points are sized relative to the mean number of occurrences in each time bin, shown (rounded to the nearest whole number) in the legend to the right of each time slice. The colored polygons at the bottom of the plot are convex hulls enclosing the taxa present at each time bin, labeled in the corresponding colors.

Figure 3.1: (continued)



We note that it ought to be possible (and may be interesting) to formulate a metric describing the evenness of morphospace occupation, a morphological equivalent of the concept of taxonomic evenness. This could be formulated, for example, by analogy to the E_{SS} metric (Peters 2004), or by comparison to stochastic simulation of random partitioning of occurrences in morphospace.

In the following section, we address the question of whether sampling differences might account for the observed changes in metrics of morphological disparity (Fig. 2.9) by applying sampling standardization methods to the diatom morphospace.

3.3.2 Subsampling of Morphospace

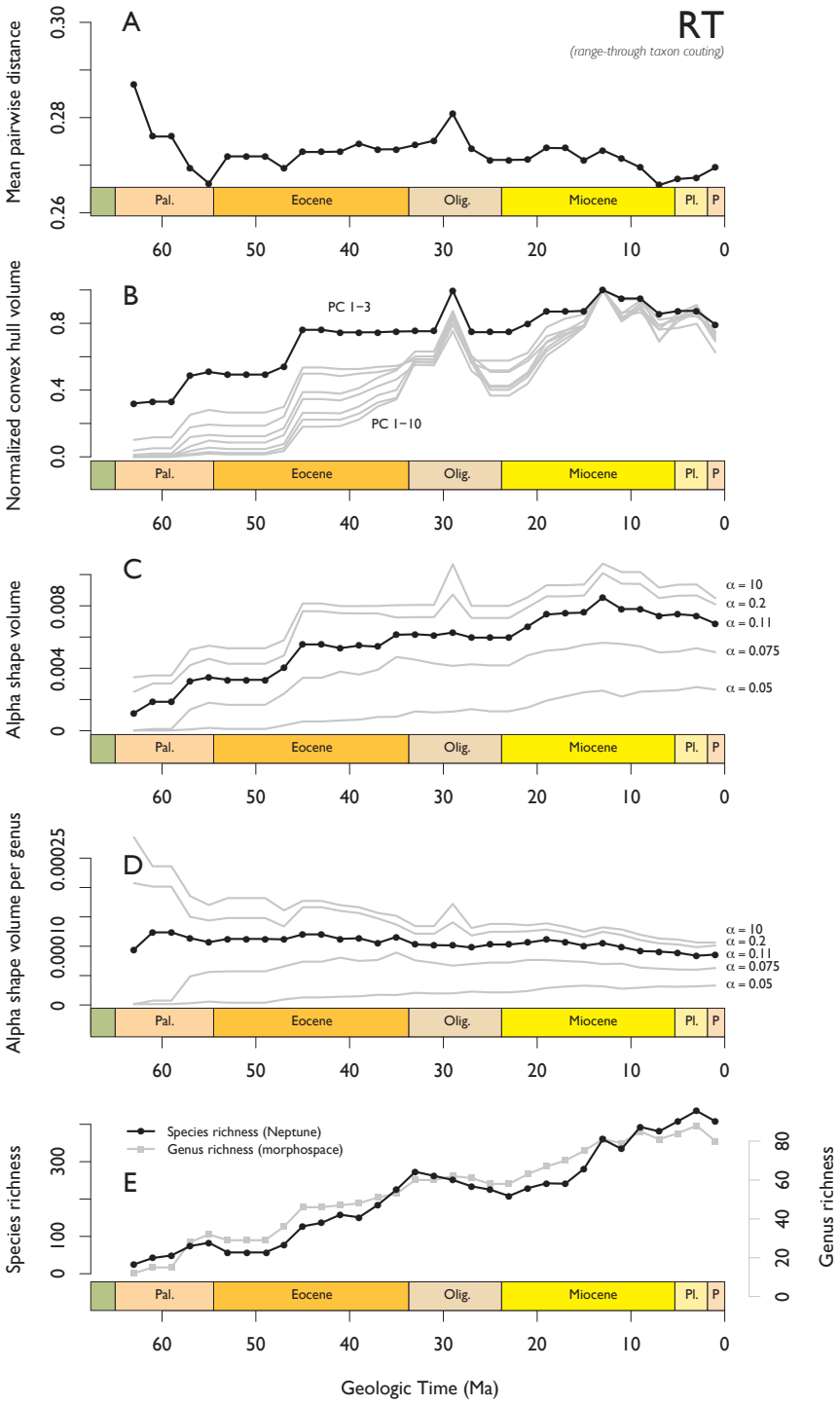
Taxon Counting

Before delving into subsampling, it is worth considering how we construct a list of the taxa present in a time bin from raw data of fossil occurrences. Curves of diatom taxonomic diversity have conventionally been compiled using the range-through method of taxon counting (Spencer-Cervato 1999; Rabosky and Sorhannus 2009), in which a taxon is counted as present in any time bin between its first and last appearance, regardless of whether or not it is actually observed in that time bin. Metrics of morphological disparity and taxonomic diversity for the Cenozoic diatom morphospace under the range-through method of taxon counting are shown in Figure 3.2.

Range-through (RT) taxon counting is intuitively appealing, because we know taxa must have been extant between their first and last appearance. However, this method has fallen out of favor because it has been shown to suffer from a number of significant biases (such as the Signor-Lipps effect and other edge effects) that distort the form of the resulting diversity curve (reviewed in Alroy 2010a). An alternative method counts only those taxa actually observed in a time bin (sampled in-bin, SIB). Although SIB taxon counting underestimates standing diversity in time bins with poor sampling, it is immune to most of the other biases affecting the RT method, and is thus generally the preferred method of taxon

Figure 3.2 *(following page)*: Metrics of morphological disparity (A-D) and taxonomic diversity (E) for the Cenozoic morphospace of marine planktonic diatoms, populated using range-through (RT) taxon counting of *Neptune* database occurrences. A, Mean pairwise distance between genera, (character state mismatches over possible matches). B, Convex hull volume containing genera, normalized to largest value; black line is volume calculated over the first three PCO axes, grey lines are volume over the first four, five, etc. up to ten PCO axes. C, Alpha shape volume containing genera; black line is volume for α-value chosen by inspection to best capture occupied volume across time bins, grey lines are other α-values. $\alpha = 10$ recovers the convex hull solution. D, Alpha shape volume (as in C) divided by number of genera. E, Species-level diversity from the *Neptune* database (includes taxa omitted from morphospace analysis) in black; genus-level diversity in morphospace analysis in grey.

Figure 3.2: (continued)



counting (Alroy 2010a). The disparity and diversity metrics for the Cenozoic diatom morphospace using SIB taxon counting are shown in Figure 3.3.

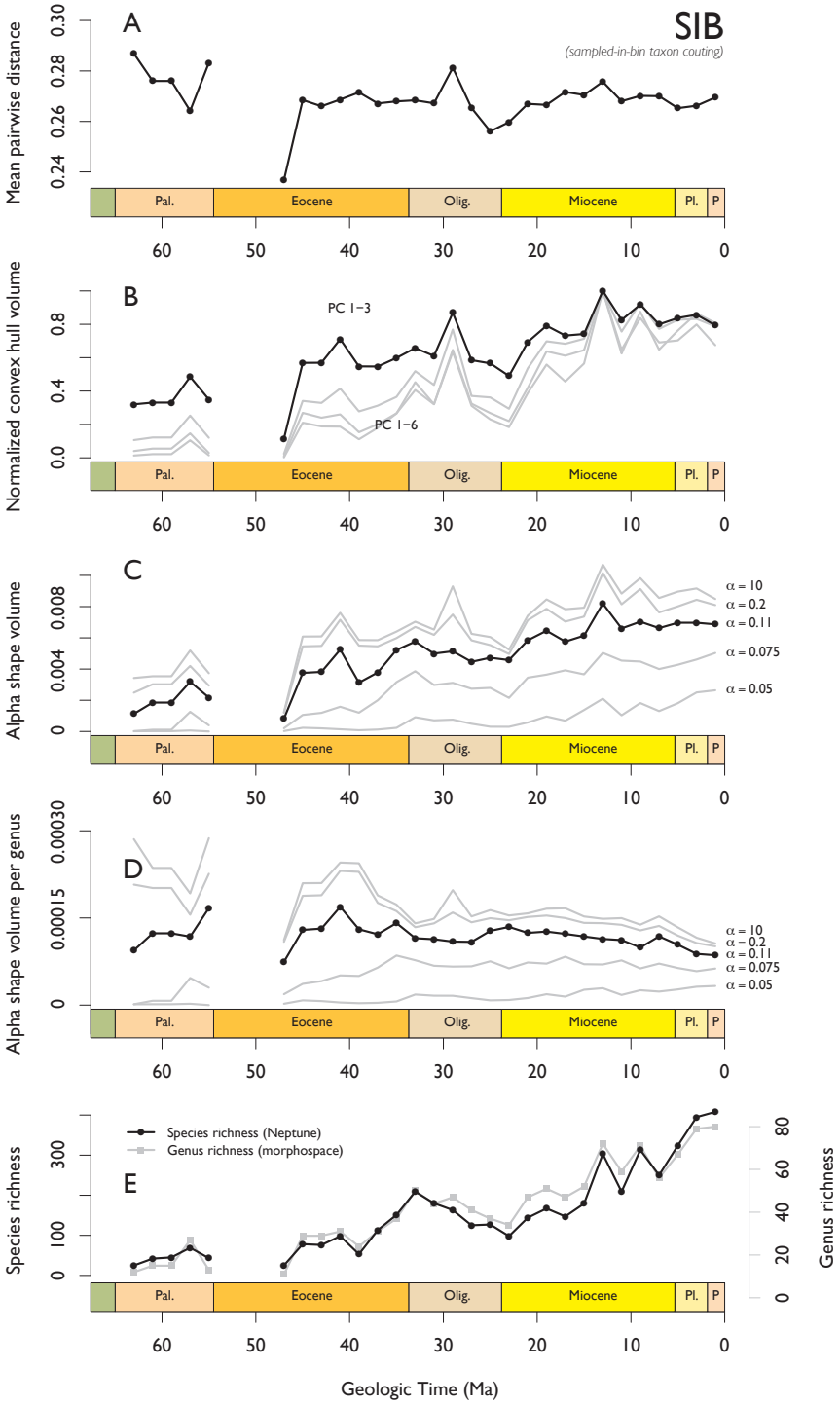
Comparing the disparity metrics calculated under RT (Fig. 3.2) and SIB (Fig. 3.3) illustrates that the method of taxon counting does not affect the first-order patterns observed. In both figures, metrics of the separation between taxa in morphospace (A, D) are approximately stationary through time, while metrics of the total volume of morphospace occupied (B, C) show an increase with time. However, the curves drawn under SIB are noisier, while the RT curves are smoother, reflecting bin-to-bin differences in sampling (with intervals of poor sampling, perhaps due to differences in preservation, masked by the RT method). Besides the obvious sampling gap in the earliest Eocene, for which no diatom data are present in the *Neptune* database, these "dips" in the SIB curves (relative to the RT curve) also highlight the potential of sampling bias to influence the disparity metrics. The dips at 47 Ma and 39 Ma in the SIB diversity curve, for example, have corresponding dips in the convex hull and alpha shape volume curves, but these dips are absent in the RT curves. Since we thus know these dips are due to sampling (taxa not counted but known to have existed), this further reinforces the need to correct for sampling before interpreting disparity metrics, particularly those describing the volume of occupied morphospace.

Uniform Subsampling

Under CR subsampling (Fig. 3.4), measures of taxonomic diversity and some measures of morphological disparity show different temporal trajectories than under SIB (Fig. 3.3). Rabosky and Sorhannus (2009) described Cenozoic diatom diversity under various methods of subsampling in detail, so we go no further here than to confirm that our results (Fig. 3.4E) agree: we find a much-attenuated, roughly twofold rise in diversity, compared to the fourfold rise under SIB (Fig. 3.3E), over the course of the Cenozoic Era. Peak diversity under CR is reached in the early Oligocene (rather than in the Pleistocene under SIB), with a more pronounced Oligocene diversity crash and a subsequent recovery to

Figure 3.3 ***(following page)*****:** Metrics of morphological disparity (A-D) and taxonomic diversity (E) for the Cenozoic morphospace of marine planktonic diatoms, populated using sampled-in bin (SIB) taxon counting of *Neptune* database occurrences. Metrics as explained in Fig. 3.2.

Figure 3.3: (continued)



early Oligocene levels diversity through the remainder of the Cenozoic Era.

The metrics of morphological disparity describing the distance separating taxa in morphospace show much the same trajectory under CR (Fig. 3.4A and D) as under SIB (Fig. 3.3A and D). The per-genus volume of morphospace occupied (Fig. 3.4D) shows a stationary pattern through time, much as under SIB (Fig. 3.3D). Similarly, mean pairwise distance (Fig. 3.4A) shows a broadly stationary pattern, much as under SIB (Fig. 3.3A), albeit with a less pronounced peak in the mid-Oligocene and a more accentuated Oligocene-Miocene trough.

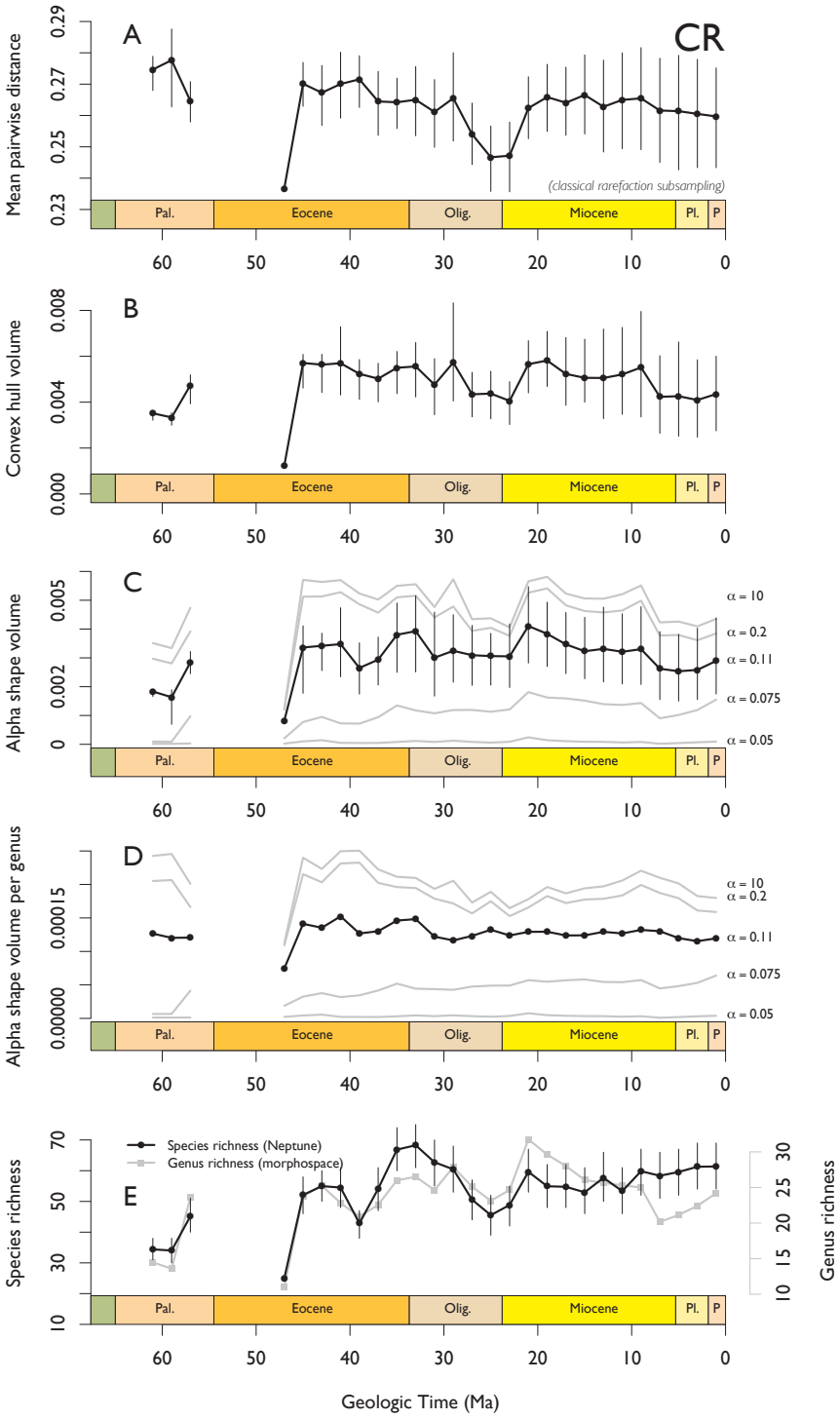
In contrast, those metrics of disparity describing the total volume of morphospace occupied show results under CR (Fig. 3.4B and C) that are qualitatively different than under SIB (Fig. 3.3B and C). Both convex hull volume (Fig. 3.4B) and alpha shape volume (Fig. 3.4C) show a broadly stationary trajectory under CR, compared to the twofold increase under SIB. Although there is an increase in occupied volume from the Paleocene to the Eocene in both the CR and SIB results, the subsequent trajectory is flat under CR where there is an increase under SIB. The spikes in occupied volume at the 41, 29, and 12 Ma time bins are attenuated under CR, perhaps because taxa responsible for an expansion of occupied space, located at the extremes of morphospace, are sampled only in some of the subsampling iterations.

The results for UW and OW subsampling are very similar to those for CR. The results for these analyses are thus provided in Appendix A (Figs. A.1 and A.2).

In summary, all disparity metrics show broadly stationary patterns when based on *Neptune* occurrence data subsampled to a uniform sampling level. Those disparity metrics describing the separation among taxa in morphospace (mean pairwise distance and mean alpha shape volume occupied per list) do not change substantially compared to the raw (SIB) results, while those metrics describing the volume of morphospace occupied (by convex hull and alpha shape) lose the increasing trend seen under SIB when subsampled.

Figure 3.4 ***(following page)*****:** Metrics of morphological disparity (A-D) and taxonomic diversity (E) for the Cenozoic morphospace of marine planktonic diatoms, populated using *Neptune* database occurrences subsampled to a quota of 100 occurrences by classical rarefaction with 10,000 iterations. Metrics as explained in Fig. 3.2; error bars show 95% confidence intervals of subsampling. Error bars omitted from genus diversity curve for clarity.

Figure 3.4: (continued)



Subsampling by SQS

Although SQS is conceptually distinct from the uniform item quota subsampling methods (CR, UW, and OW), the morphospace metrics calculated under our version of SQS (Fig. 3.5) are similar to those obtained through the other methods (Figs. 3.4, A.1, and A.2). Under SQS, mean pairwise distance (Fig. 3.5A) shows a generally stationary pattern (again with a very slight net decline representing at most a few percentage points in dissimilarity), much as in the other analyses.

Convex hull volume also shows a generally stationary pattern under SQS (Fig. 3.5B), albeit with slightly more variability than under CR. Alpha shape volume through time (Fig. 3.5C) also shows greater amplitude variability under SQS than CR, and although the net increase over the Cenozoic is still far less than under SIB, there is a clearer increase under SQS than under CR. However, this increase may be an artifact of the choice of α parameter, which was chosen at $\alpha = 0.11$ to optimally describe the arrangement of taxa in the raw dataset and may not adequately capture morphospace occupancy of smaller subsamples with a different arrangement of taxa. Indeed, volumes calculated with higher values of α (0.2 and 10, upper grey curves in Fig. 3.5D) show a more stationary pattern.

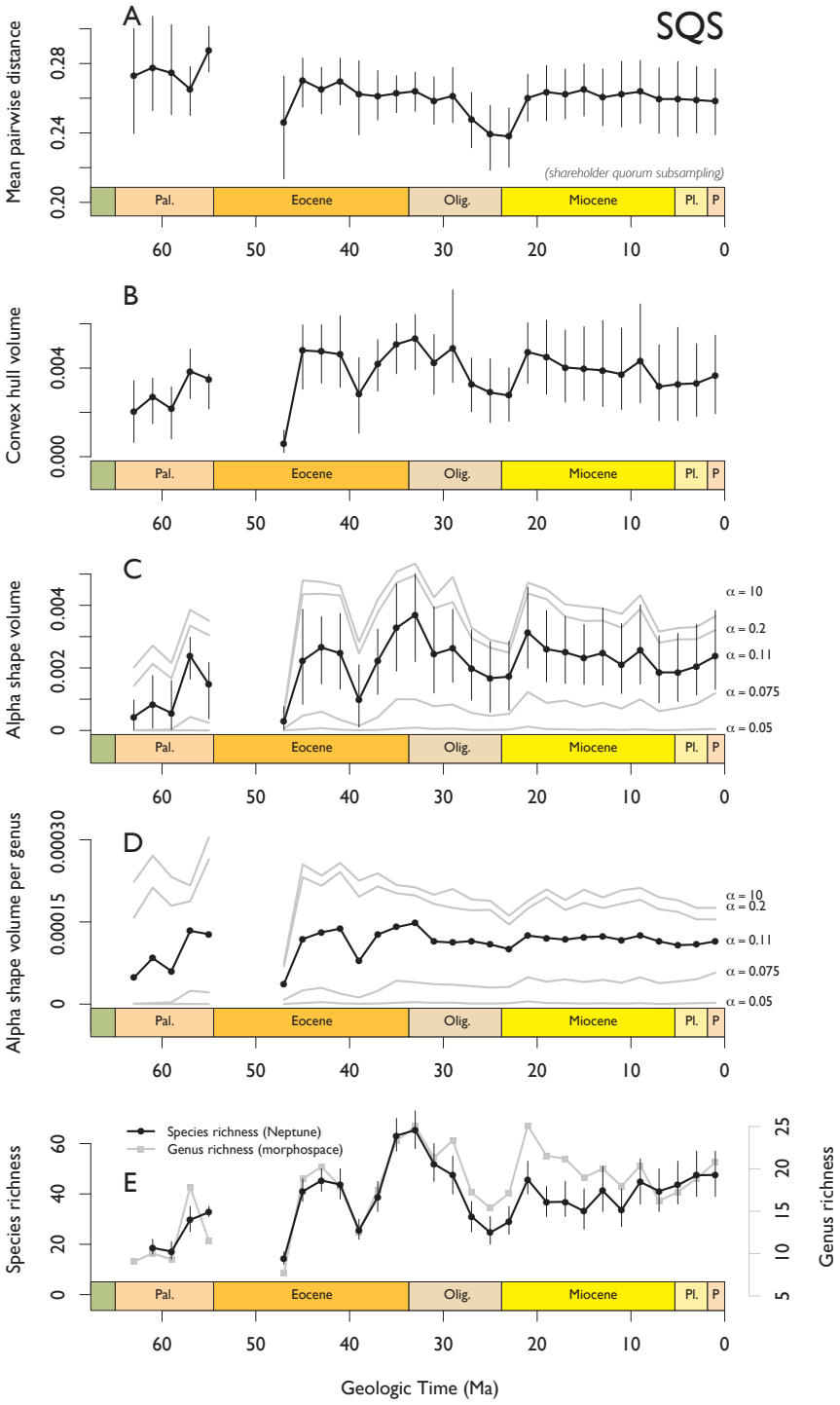
Finally, per-genus alpha shape volume (Fig. 3.5D) shows a stationary pattern over much of the Cenozoic under SQS, similar to the results under CR, though volumes in the Paleocene time bins and one Eocene time bin are lower under SQS than under CR, which suggests a slight increase over time.

The Cenozoic trajectory of taxonomic diversity (Fig. 3.5E) is greatly flattened, much as in the uniform subsampling method results. However, the Eocene-Oligocene peak in diversity under SQS greatly exceeds the diversity recovered subsequent to the Oligocene. In this regard, the SQS diversity curve resembles the O2W curve presented by Rabosky and Sorhannus (2009).

It should be noted that the diatom diversity curves obtained by subsampling methods have not been universally accepted by micropaleontologists (Lazarus et al. 2012a). A criticism of these methods, including SQS, is that they can perform poorly under changes in relative frequency distributions. In essence, if

Figure 3.5 ***(following page)*****:** Metrics of morphological disparity (A-D) and taxonomic diversity (E) for the Cenozoic morphospace of marine planktonic diatoms, populated using *Neptune* database occurrences subsampled by to a uniform coverage of 0.5 by shareholder quorum subsampling with 1,000 iterations. Metrics as explained in Fig. 3.2; error bars show 95% confidence intervals of subsampling.

Figure 3.5: (continued)



relative frequencies are very evenly distributed to begin with and become very uneven through time, subsampling could significantly underestimate diversity in the more uneven intervals. A similar concern has been raised with regard to increases in provinciality through time (changes in β diversity), and an alternative diversity curve more similar to the canonical view (Spencer-Cervato 1999) has been put forth by Lazarus et al. (2012a). It is obtained by adding empirical correction factors to subsampled diversity curves to account for changes in evenness and provinciality. These contrasting views are discussed in more detail on page 98.

Summary of Subsampling Results

The results of morphospace analyses under different subsampling methods show the following:

1. When sampling bias is corrected by randomized subsampling, all disparity metrics show stationary patterns or, at most, directional changes of small magnitude (a small decrease in mean pairwise distance in all analyses and a small increase in occupied volume under SQS).

2. Morphological diversification in Cenozoic diatoms is described as stationary once sampling differences are taken into account. This is true for both measures of average morphological distances among taxa and the total range of morphologies explored, and is in agreement with the results of the comparison of morphological with molecular and phylogenetic distance in Chapter 2 (page 35).

3. Disparity metrics describing the average dispersion of taxa in morphospace (mean pairwise distance and per-genus alpha shape volume) are less sensitive to sampling bias than those metrics describing the total extent of morphospace occupied (convex hull and alpha shape volume).

By using subsampling methods, we seek to discover something about the nature of morphological diversification by correcting for differences in sampling.

In the following section, we pursue the same goal using a different approach, by examining aspects of the data that are independent of those secular variations in sampling.

3.3.3 Occupied Morphospace Per List

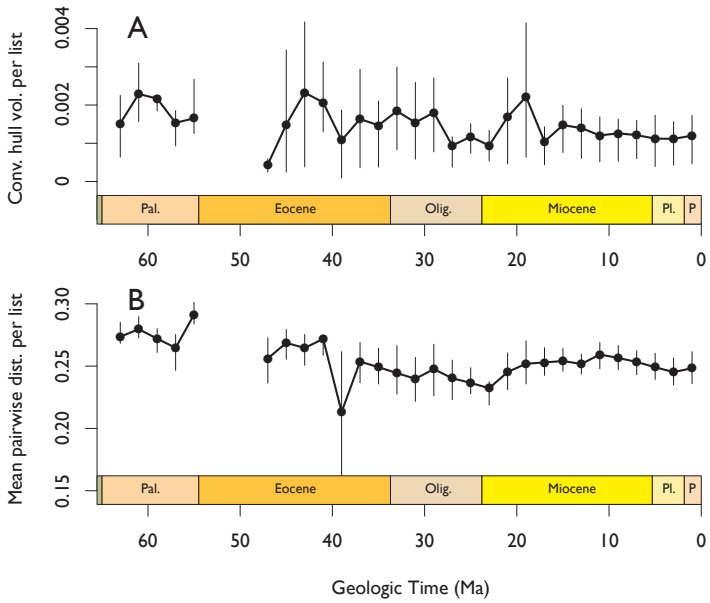


Figure 3.6: Metrics of "α disparity", the average morphological disparity represented by a taxonomic list, measured in (A) convex hull volume (in three dimensions) and (B) mean pairwise distance. Error bars show the middle 50% of values, i.e. the 25th and 75th percentiles. Note that α disparity is unrelated to the concept of alpha shapes used to quantify occupied morphospace volume.

An alternative means to overcome the problem of sampling bias is to look at measures of morphological disparity calculated for individual lists (the sets of taxa reported from a particular depth in a particular borehole). A helpful context for this approach is to consider the notions of α and β taxonomic diversity.

We can consider global taxonomic diversity, S, to consist of a local component—described by the length of a taxonomic list at a particular location, α—and a component describing how different any given list is from another. A

useful definition relating these components is that of Whittaker (1960), who defined this β component as:

$$\beta = \frac{S}{\overline{\alpha}}$$

Using this definition, we can consider β diversity as the number of potentially unique communities (or the number of nonoverlapping lists of average list length). By rearranging this expression as $S = \overline{\alpha} \cdot \beta$, it becomes clear that changes in observed global diversity can be due to either changes in the per-locality diversity or changes in the taxonomic similarity among localities (or some combination of the two). Such changes in the components of global taxonomic richness have been explored, for example, in Paleozoic marine animals by Sepkoski Jr (1988).

By analogy, we can think of a geographic structure in morphological disparity, consisting of a component describing the local morphological disparity ("α disparity") and a component describing how morphologically different communities are from one another ("β disparity"). We calculated the average α disparity for both mean pairwise distance and occupied convex hull volume as per-list disparity metrics in each time bin (Fig. 3.6).

Both average per-list convex hull volume (Fig. 3.6A) and per-list mean pairwise distance (Fig. 3.6B) show broadly stationary patterns, though the latter shows a slight decline through time (as does mean pairwise distance at the global level, Fig. 3.3A). Although methodological bias toward constant list length during data collection has been suggested for micropaleontological data (Lazarus 2011), such a bias would simply imply that these results have been standardized for secular changes in taxonomic diversity. These results are consistent with the largely stationary patterns observed at the global scale under subsampling and support an overall picture of Cenozoic stasis in diatom morphological evolution.

The per-list volume results (Fig. 3.6A) also suggest that the increase in occupied morphospace volume seen at the global scale in the raw data (Figs. 3.3B and C) must have a spatial component: if the increase in occupied morphospace

volume is not due to an increase in the volume occupied by individual lists (and, by extension, by local assemblages), it stands to reason that the increase is due to the addition of more lists occupying similar-sized but non-overlapping volumes of morphospace. As explained above, we describe this as a rise in β disparity. One might imagine that the increasing latitudinal temperature gradients observed through the Cenozoic Era (Zachos et al. 2001) might contribute to such an increase.

Though we can confidently infer this rise in β disparity in our data, we cannot determine whether it represents a true geographic differentiation in diatom disparity or whether this is an artifact of the secular increase in the number of lists sampled. Nonetheless, we can rule out Cenozoic morphological diversification at the local scale, finding instead a stationary pattern in α disparity consistent with the Cenozoic stasis in our other results.

3.3.4 Sensitivity of Results to Methodological Choices

Constructing a morphospace and using it to measure secular changes in disparity involves numerous methodological choices. We have already investigated the effect of one of these choices, the taxon counting method, in Section 3.3.2. In the following, we test the sensitivity of our disparity metrics to further important methodological choices that are commonly unexamined: how to find a low-dimensional representation of the morphospace (the choice of ordination method) and how much incomplete data to reject before constructing the morphospace (the choice of data culling threshold).

Ordination Method

In order to investigate the sensitivity of our results to the choice of ordination method, we repeated the calculation of convex hull volumes and alpha shape volumes through time using another ordination method commonly used in morphospace studies (e.g. by Huntley et al. 2006; Shen et al. 2008): non-metric multidimensional scaling (NMDS). Unlike PCO, NMDS is not an eigenvector

method; rather, a fixed number of dimensions is chosen *a priori* and the best representation of the data in those dimensions is found numerically. The method proceeds through successive iterations until an acceptable (but not necessarily unique or optimal) solution is found. We carried out this analysis using the `isoMDS()` function from the *MASS* package (Venables and Ripley 2002).

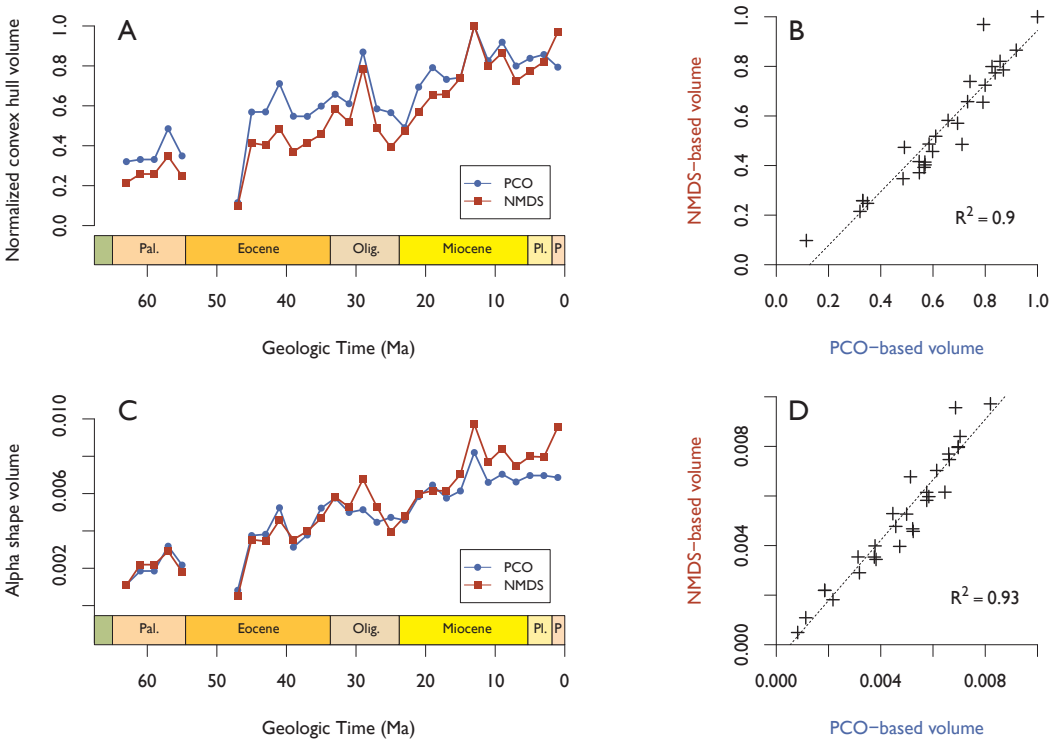


Figure 3.7: Plots illustrating the (in-)sensitivity of the volume-based disparity metrics to the choice of ordination method. A, normalized convex hull volume through time, calculated for three dimensions only, using PCO ordination (blue plot points) and NMDS (red plot points). B, crossplot of the PCO and NMDS results in (A), with linear model and squared correlation shown. C, alpha shape volume through time for both ordination methods. D, crossplot and squared correlation of results in (C).

Figures 3.7A and C show the resulting comparison of morphospace volume metrics calculated using NMDS with three dimensions specified (red points) and the first three PCO axes (blue points). The results are very similar, and when the

timeseries resulting from one ordination procedure is plotted against those resulting from the other (Figs. 3.7B and D), the closeness of this correlation can be summarized with an R^2 value (0.90 and 0.93 for convex hull and alpha shape volumes, respectively).

These results suggest that metrics of occupied morphospace volume are not sensitive to the choice of ordination method.

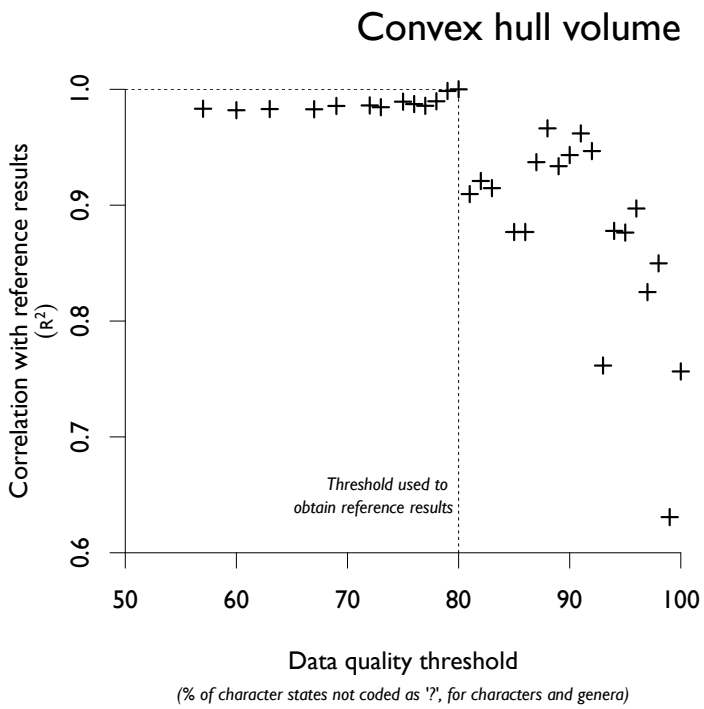
Data Culling

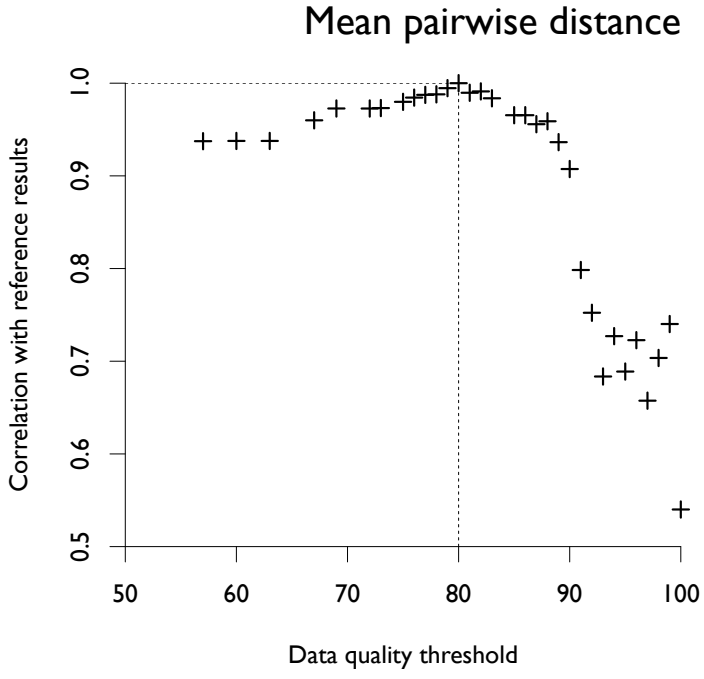
Virtually all paleontological datasets contain missing data, and this is particularly true of those used to construct morphospaces. The possible causes of missing entries in the morphospace matrix used here are discussed in more detail in Section 2.3.2, but a crucial question at the outset of a morphospace study is: how much valid data should a genus or character have to be included in the analysis? The edge cases are trivial to decide: a genus with no valid character states or a character with no valid entries for any genus adds no information and obviously ought to be excluded. Likewise, genera and characters with entirely valid entries ought to be included. Where the line is drawn in between these extremes is to some extent an arbitrary decision; in this study, we chose a threshold of 80% completeness.

In order to investigate the sensitivity of our results to different choices of data quality threshold, we repeated our analysis under the entire range of completenesses represented in our data, ranging from including all the data collected at one extreme (a threshold of 57% or more of observed states) to including only complete genera and characters at the other (a threshold of 100% observed states). The data culling algorithm was applied in the same manner as for Chapter 2, removing first characters and then genera until both reached the desired threshold of data quality. As before, we only considered unobserved entries in calculating completeness, since we consider the other types of missing data (the cases where states are either inapplicable to a taxon or where multiple states apply) to constitute important information.

Figure 3.8 ***(following page)*****:** Plots showing the sensitivity of disparity metrics to the quality threshold required for data included in the analysis. In both plots, each plot point represents a comparison between the results reported in Chapter 2 (the "reference results") and the results of an analysis with data satisfying a certain level of completeness, expressed as a correlation coefficient (R^2) between the two sets of results. The plot above shows results for a metric of the total extent of occupied morphospace (convex hull volume), the plot below shows results for the dispersion metric (mean pairwise distance). Because the reference analysis used an 80% completeness threshold, the correlation is perfect at that threshold (the method of taxon counting in all cases was SIB).

Figure 3.8: (continued)





We compared the convex hull volume and mean pairwise distance results obtained under each data quality threshold to the "reference results" under the 80% threshold presented in Chapter 2. Rather than plotting the timeseries for each of these comparisons (like in Figs. 3.7A and C), we summarized each comparison using the R^2 correlation coefficient (like in Figs. 3.7B and D).

The R^2 values summarizing the comparison of analyses under each data quality threshold with the reference results are shown in Figure 3.8. These results show that neither convex hull volume (Fig. 3.8A) nor mean pairwise distance (Fig. 3.8B) are sensitive to the addition of more data of lower quality. Even setting the most permissive threshold (including all the data collected) yields time series that are highly correlated ($R^2 > 0.9$) with the reference results. The results also remain correlated above ~0.9 as data are removed until the data quality threshold exceeds about 90% completeness, beyond which correlations decline. Results under the most stringent data quality threshold (100% complete characters and genera only) show relatively weak correlations of only 0.5–0.6. Mean pairwise distance appears to be more sensitive than convex hull volume to changes in data quality threshold.

In order to clarify whether the results with R^2 values suggesting weak correlations with the reference results are in fact qualitatively different, we plotted a comparison between the results using the most stringent data quality threshold (100% completeness, with only 32 characters retained) and the reference results (80% completeness), shown in Figure 3.9. In spite of the low R^2 values, the results are qualitatively similar. Convex hull volume increases in both cases (Fig. 3.9A) while mean pairwise distance remains roughly constant in both cases (Fig. 3.9C), although the absolute values of distance are lower under the more stringent threshold.

3.3.5 Testing Character Sets for Specific Evolutionary Hypotheses

We have thus far approached our goal, to make biological inferences about the morphological evolution of the diatom frustule, by summarizing morphological

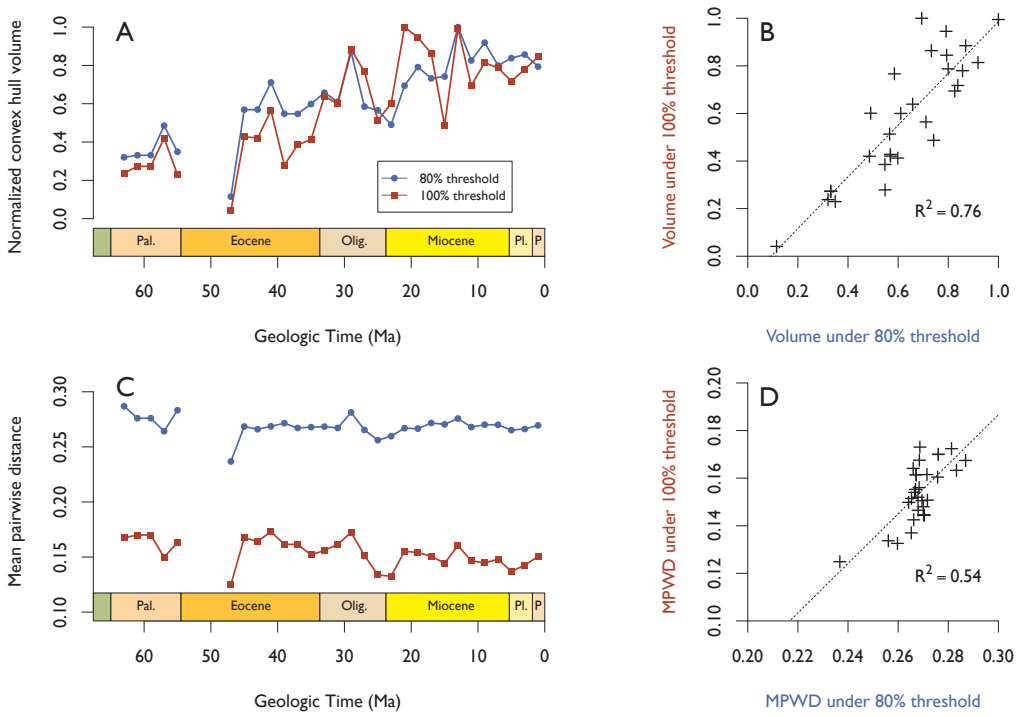


Figure 3.9: Comparison of results under two different thresholds of data quality, 80% (as used in the results above and in Chapter 2, shown in blue) and 100% observed character states (no "?" entries in the morphospace matrix, shown in red). A, normalized convex hull volume through time, calculated for three dimensions only, using PCO ordination. B, crossplot of the results in (A), with linear model and squared correlation shown. C, alpha shape volume through time under both data quality thresholds. D, crossplot and squared correlation of results in (C). MPWD stands for mean pairwise distance.

data and abstracting it through metrics of disparity (Chapter 2) and correcting those measures for sampling differences (this chapter). These results all seem to point towards Cenozoic stasis.

The morphological data set underlying the morphospace analysis also permits an analysis of morphological evolution from a fundamentally different approach, if we momentarily set aside concerns about sampling. Rather than examining the data abstractly and in aggregate, we can analyze the morphological data directly to examine how the prevalence of taxa with different sets of morphological characters has changed through time. A similar approach has previously been used to categorize Phanerozoic animals by anatomical and ecological traits to document major shifts in the proportions of, for example, physiologically unbuffered to physiologically buffered taxa, or predator to non-predator taxa (Bambach et al. 2002). These categories were associated with evolutionary hypotheses about mass-extinction kill mechanisms (Knoll et al. 1996) and ecological escalation (Vermeij 1987), respectively.

By analogy, we can parse our morphological data *a priori* by criteria related to hypothesized drivers of diatom evolution. For example, predation has been suggested to play an important role in diatom evolution (Smetacek 2001; Hamm and Smetacek 2007) and we can identify characters that might relate to defense against predation, like spines and projections or ribs and costae buttressing and strengthening the valve. Then, we can investigate whether the prevalence of these characters changed through time, as would be expected under the hypothesized selective pressure. If we were able to detect systematic changes in the proportion of character states expected under a given scenario, we might question the picture of stasis painted by the subsampling exercises and the alpha disparity results above.

We assembled lists of characters expected to change under changes in four factors that have been identified as central to Cenozoic evolution in diatoms: predation, sinking (Raven and Waite 2004), viral attack (Smetacek 1999), and silica availability (Finkel and Kotrc 2010). For each chosen relevant character, we sorted character states into favorable and unfavorable categories (e.g. for

predation, character states indicating possession of spines were assigned to the favorable category, those states indicating absence of spines to the unfavorable category). The complete listing of characters and assigned states are tabulated in Appendix E.

The results (Fig. 3.10) show a remarkable absence of trends through time. The proportion of morphological character states thought to be associated with specific hypothesized drivers of evolution in diatoms are essentially constant through Cenozoic time. These results portray stasis in morphological evolution that is consistent with the other lines of evidence presented here. However, we note that the absence of trends in these characters do not necessarily imply a lack of response to these selective pressures, since some responses may simply not be visible in our data. For example, our morphospace does not capture changes in cell size, although this may be an important factor in mechanical strength and thus predation resistance (Hamm et al. 2003), and (Finkel et al. 2005) documented a Cenozoic decrease in the size of diatom frustules that may point to just such a response.

3.4 Conclusions

The substantial Cenozoic rise in sampling through time calls into question the marine planktonic diatom disparity results presented in Chapter 2, which show a rise in occupied morphospace volume, in contrast to the stasis seen in all other metrics. Two further analyses presented here highlight the need to take sampling differences into account before interpreting disparity metrics. First, the differences between volume-based disparity metrics calculated under different methods of taxon sampling (SIB and RT) suggest that these metrics are affected by sampling. Second, illustrating the number of occurrences represented by each taxon in a morphospace plot shows that morphospace is occupied unevenly and raises the possibility that less-intensive sampling of more recent time bins may have led to lower reported volumes of morphospace occupation.

The plotting of morphospace occupation "density" permitted by the use of an

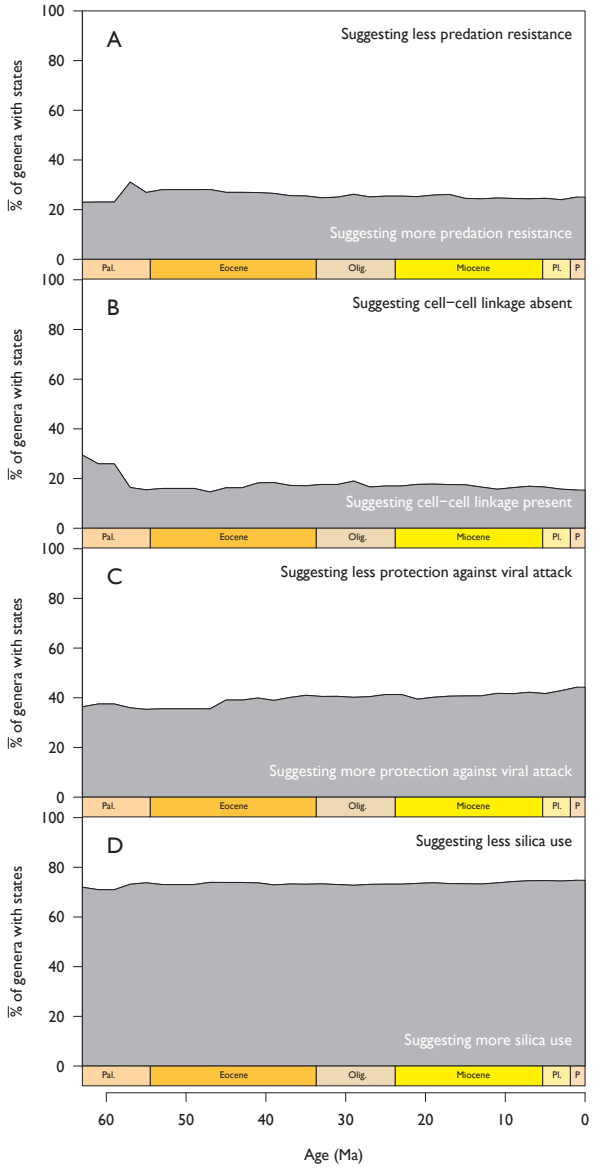


Figure 3.10: Prevalence through time of sets of characters expected to change under different hypothetical Cenozoic drivers of diatom evolution. A, characters related to predation resistance; B, characters indicating cell-cell linkage, thought to impact sinking rates; C, characters thought to confer resistance against viral attack; D, characters impacting silica use.

occurrence-based database leads us to formulate a notion of "morphological evenness." Analogous to taxonomic evenness, which describes the distribution of individuals (or, in paleontological studies, occurrences) among taxa, morphological evenness would describe the distribution of individuals (or occurrences) in morphospace. Any given abundance distribution could be, at one extreme, randomly distributed throughout morphospace; at the other extreme, occurrences could be preferentially concentrated in one area. We suggest that quantifying this notion would be an interesting target for future work.

In order to address the potential sampling bias identified in these ways, we recalculate the disparity metrics presented in Chapter 2 under various methods of subsampling. We find that, under subsampling, the increases in occupied morphospace volume seen in Chapter 2 largely disappear, and all disparity metrics show essentially stationary results. These results suggest morphological stasis through Cenozoic time when sampling differences are corrected in this fashion.

Comparing the disparity metrics calculated under subsampling to those calculated from the data directly suggests that the metrics describing the volume of occupied morphospace are more sensitive to sampling differences than those describing the distances among taxa (or, put another way, their dispersion in morphospace). These results agree with the findings of Butler et al. (2012) and Ciampaglio et al. (2001), albeit using different metrics.

In seeking a direct measure of disparity insensitive to sampling intensity, we introduce the concept of a geographic component to morphological disparity. By analogy to Whitaker's (1960) α and β components of taxonomic diversity, we suggest the notions of α and β disparity. We find that $\bar{\alpha}$ disparity (as quantified by the mean of either convex hull volumes or the mean pairwise distances across lists) remains roughly constant through time. These results support Cenozoic stasis in diatom morphological evolution.

Constant $\bar{\alpha}$ disparity through time is compatible with the observations of roughly constant total disparity under subsampling. If the subsampling results were to be rejected in favor of the results in Chapter 2, however (see caveats

below), constant $\bar{\alpha}$ disparity would imply that the rise in total disparity resulted from an increase in β disparity.

As a by-product of applying subsampling methods to diatom morphospace, we present a taxonomic diversity curve of diatoms under SQS based on the *Neptune* database. We find results similar to other subsampling methods, with a flattened diversity curve showing peak diversity near the Eocene/Oligocene boundary and a pronounced Oligocene decline in diversity. In the SQS diversity curve, this Eocene/Oligocene peak far exceeds the species richness recovered subsequently, and is thus most similar to the O2W results reported by Rabosky and Sorhannus (2009).

The diatom diversity curves obtained by subsampling methods, however, have not been universally accepted by micropaleontologists (Lazarus et al. 2012a), because they can perform poorly under changes in relative frequency distributions. Though this issue is discussed in more detail on page 98, it is sufficient here to point out that the stationary results of the volume-based disparity metrics and the sampling-corrected diversity curves are dependent on whether subsampling methods are believed to provide a more accurate view than the raw data, or whether they simply trade on bias for another. The other results indicating stasis, however—the distance-based disparity metrics, the disparity metrics per-list, and the comparison of morphospace to molecular phylogeny—do not depend on subsampling.

In a sensitivity test comparing our morphospace volume results using PCO to those using NMDS, a substantively different, non-eigenvector ordination method, we find similar results in both and conclude that our results are not sensitive to ordination method. In a similar sensitivity test repeating our analyses after culling more or less of the data by completeness, we find that our results are also robust to choices in data quality threshold.

In summary, when sampling biases are taken into account using subsampling methods as well as sampling-independent metrics of disparity, our results point toward Cenozoic stasis in the occupation of planktonic diatom morphospace. This suggests diatoms had reached peak disparity by the early Cenozoic Era,

while taxonomic diversity continued to rise, albeit more gradually than the canonical diversity curve would suggest. Though we have not analyzed diversity and disparity from the origin of the clade, our results point to a decoupling of taxonomic and morphological diversification akin to the "asymmetric diversification" reported for many other groups.

More broadly, these results make clear that a complete view encompassing all aspects of morphological disparity must consider sampling biases. The use of occurrence-based databases to populate morphospaces allows these biases to be addressed using well-established, quantitative methods.

4

Morphospaces and Databases: The Evolution of Diatom Shape and Diversity through Time

Abstract

The diversity of diatom form inspired Art Nouveau designers, an interest renewed by recent advances in biomimetic design. The fossil record provides two windows on the diversification history of diatoms: taxonomic diversity and morphological disparity. Conventionally, marine planktonic diatom diversity describes a steep, almost monotonic rise through Cenozoic time. Subsampling methods used to address the associated rise in sampling reveal a more stationary pattern, with peak diversity in the mid-Cenozoic, whether by established methods or a new method

(shareholder quorum subsampling, SQS). However, these methods may underestimate diversification if evenness decreases. In order to measure morphological disparity, we constructed an empirical morphospace based on discrete characters. Mean pairwise distance, a disparity metric describing the density of taxa in morphospace, shows little secular change , while convex hull volume, a measure of the extent of occupied morphospace, increases through time. Since we populated the morphospace with occurrence-based data, we can apply subsampling algorithms to these disparity metrics. Mean pairwise distance is largely unaffected, while the increase in occupied volume largely disappears under subsampling. Depending on the metric used, characterizing diatom diversification thus depends upon whether a literal reading of the fossil record or the use of subsampling algorithms is preferred. While this may prompt a reexamination of evolutionary narratives prominently featuring diatom diversification, changes in abundance and silicification may also affect the diatoms' biogeochemical importance. For biologically inspired design, an early exploration of diatom morphospace suggests that fossil forms should be considered alongside extant diatoms.

4.1 Introduction

The diversity of diatom form has been a source of fascination and inspiration since diatom frustules were first described by the 19th Century pioneers of micropaleontology (Ehrenberg 1838; Haeckel 1904) and their shapes applied to Art Nouveau architecture and design, like René Binet's design for the Printemps department store (Proctor 2006) or Hendrik Petrus Berlage's jewelry imitating chain-forming diatoms (Netherlands Architecture Institute 2012). Many thousands of extant diatom species have been described (Mann and Droop 1996), their shapes representing a wide range of variations on a basic pill-box Bauplan—from circles to triangles, needles, and curves—with staggering variety in the geometrically arranged, hierarchical pore structure (Round et al. 1990), lending an aesthetic that evidently appealed to turn-of-the-century designers. With biomimetic design advancing from superficial aesthetic inspiration to an application of underlying structural and evolutionary principles, renewed interest

in diatoms warrants efforts toward a deeper understanding of their diversification, a cardinal feature of any clade's evolutionary history.

The fossil record provides two windows on clade diversification history: taxonomic diversity and morphological disparity. The former, often referred to as taxonomic richness or simply as diversity, is the familiar measure that tallies numbers of taxa (commonly species). The latter, disparity for short, describes the variety of shapes or the "within-group variance of form" (Erwin 2007) by directly quantifying organismal morphology. In a sense, diversity and disparity both measure variety of form, because fossil taxonomy is, of course, itself based on morphology. It is also intuitive, however, that the two approaches measure that variety in very different ways. As an extreme example, a collection containing one species of fish, one species of elephant, and one species of insect has the same taxonomic diversity as a collection of three fish species, though the former clearly represents much greater morphological disparity.

We have two tie points on both the taxonomic and the morphological diversification of diatoms—their origin and their present diversity and disparity—from which we can trivially infer a net increase through time. But the more interesting questions about what happened in between are less trivial. What were the trajectories of diatom diversity and disparity through time? Has there been a monotonic increase, or was an early rise followed by stasis or even decline? Did diversity and disparity vary in lockstep or independently?

While the fossil record of diatoms extends back to at least the early Cretaceous Period (Gersonde and Harwood 1990) and includes many occurrences from nonmarine environments, the most robust and abundant data come from deep-sea sediments of the Cenozoic Era. Although it does not represent the entire clade's evolutionary history, we focus on this record here because it allows us to consider the biases that uneven sampling through time may impart to our view of evolutionary history and process.

4.2 Reconstructing Taxonomic Diversity

Conventionally, the Cenozoic history of marine planktonic diatom diversity describes a steep, almost monotonic rise of about an order of magnitude (Spencer-Cervato 1999). This view plays a central role in a number of evolutionary narratives involving the diatoms, including their coevolution with grasses (Falkowski et al. 2004) and whales (Marx and Uhen 2010), their role in reshaping the silica cycle, and its effect on radiolarians (Lazarus et al. 2009). Although widely accepted, this view has recently been challenged (Rabosky and Sorhannus 2009). The conventional diversity curve is generated from *Neptune*, a large database of marine microfossil occurrences reported from the Deep Sea Drilling Program and Ocean Drilling Program, representing several decades of micropaleontological research (Lazarus 1994; Spencer-Cervato 1999). The diversity history derived from these occurrences is not, however, a unique result, since different methodological choices can be made in taxon counting, dealing with data imperfections, and accommodating secular variations in sampling intensity. Each of these can change the diversity curve generated.

4.2.1 Taxon counting

In order to get from database to diversity curve, occurrence data need to be divided into time bins (in our examples below, of two million year duration) and the number of taxa in each bin counted. Traditionally, this has been done by counting taxa as present in all time bins between their earliest and latest occurrences, then tallying taxa known to have existed in each time bin regardless of whether observed or inferred. This "range-through" method (RT, see Fig. 4.1) has an advantage over simply counting taxa observed, or sampled in-bin (SIB, see Fig. 4.1), because it takes into account variations in preservation and sampling from one bin to another. For example, in Figure 4.1, the plot of diatom diversity from *Neptune* using RT taxon counting makes up for missing data between 54–48 Ma; diversity during that interval was clearly not zero, as a literal reading of the SIB curve might imply.

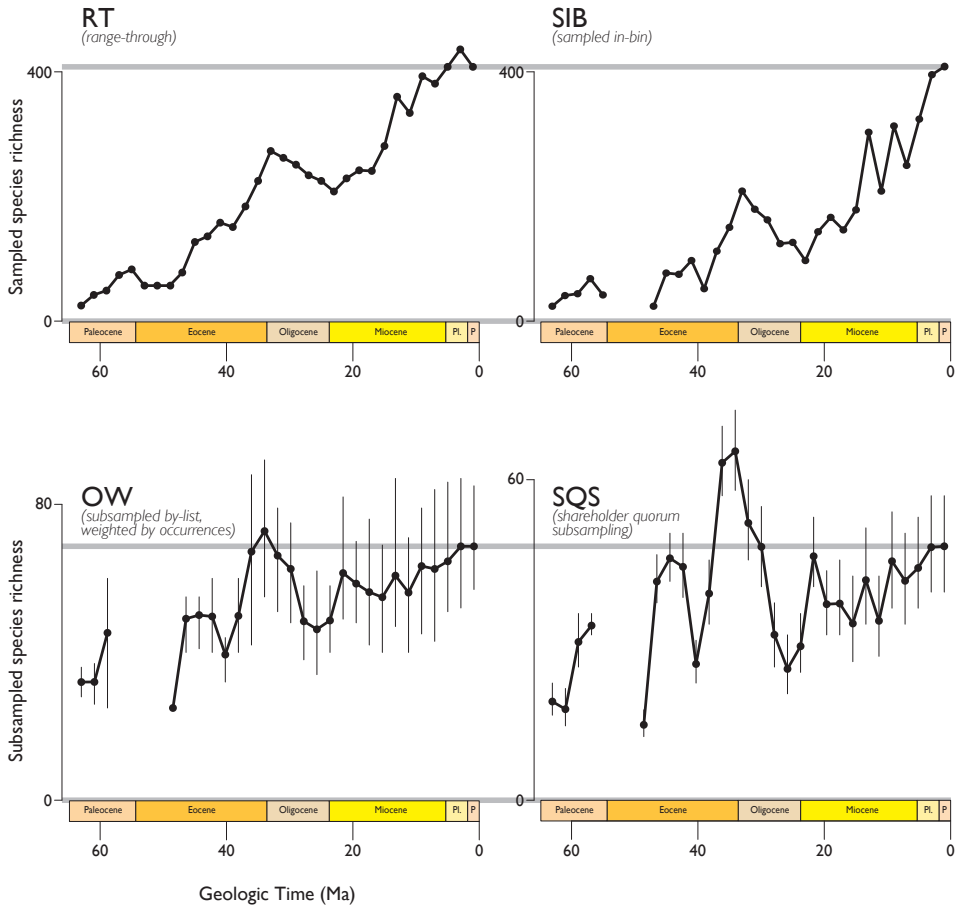


Figure 4.1: Diatom diversity through time over the Cenozoic Era, based on analysis of the *Neptune* database by range-through taxon counting (RT), sampled-in-bin taxon counting (SIB), subsampling by-lists, weighted by occurrences (OW), and shareholder quorum subsampling (SQS).

Despite its advantage in accounting for 'Lazarus taxa' (taxa that appear to go extinct, only to reappear later in the record; Flessa and Jablonski 1983), however, the RT method has fallen out of favor among many paleobiologists because it imparts a number of undesirable and potentially severe biases: the Signor-Lipps effect, the Pull of the Recent, and other edge effects reviewed by Alroy (2010a). The SIB sampling method is preferred over RT and other methods, such as tallying only those taxa known to cross the boundary between adjacent time bins, because it is immune to these biases, and the bin-to-bin sampling differences that remain can be counteracted with corrections like the part-timer sampling probability, which effectively performs a temporally localized range-through among adjacent time bins (Alroy et al. 2008). While the diversity curve for the *Neptune* diatom data obtained by SIB taxon counting does differ from the conventional curve obtained using RT in the details, the curves are rather similar in shape to first order.

4.2.2 Data quality

Generally, paleontologists worry that the fossil record underestimates the true ranges of taxa (e.g. Marshall 1990), but the record of marine microfossils is so unusually rich that the opposite has been suggested for the *Neptune* database. Marine microfossils can appear outside of their true range due to "RATs", that is, because of the physical **r**eworking of sediments (erosion and redeposition in a stratigraphically younger position), errors in the **a**ge model assigning a fossil occurrence to the wrong time bin, or **t**axonomic error (Lazarus 2011). For the curve that has become the canonical depiction of diatom diversity, these problems were addressed by manually excluding *Neptune* occurrences considered unreliable, including occurrences near depositional hiatuses (Spencer-Cervato 1999). The effect of all but the most severe instances of reworking can be obviated by setting sufficiently wide time bins, and misplaced occurrences far from the true range of a taxon are much less of an issue for SIB than RT taxon counting. Nonetheless, outliers could also be identified for removal by applying hat-shaped

models of the rise and fall in occurrences through a taxon's range (Liow and Stenseth 2007; Liow et al. 2010), but a much simpler method recently proposed just trims a certain calibrated percentage of occurrences from the beginning and end of a taxon's range—aptly named Pacman profiling (Lazarus et al. 2012b).

4.2.3 Sampling biases

An arguably graver concern for the accurate reconstruction of diatom paleodiversity is that the amount of data in the *Neptune* database greatly increases with time, as shown in Figure 4.2. This is worrisome because it is easy to imagine a situation where true diatom diversity in fact remained constant, but a steadily increasing number of samples through time captures more species from younger intervals, giving a false impression of rising diversity. Although this concern was noted in the first explorations of the *Neptune* dataset (Spencer-Cervato 1999), it was only recently addressed in detail (Rabosky and Sorhannus 2009).

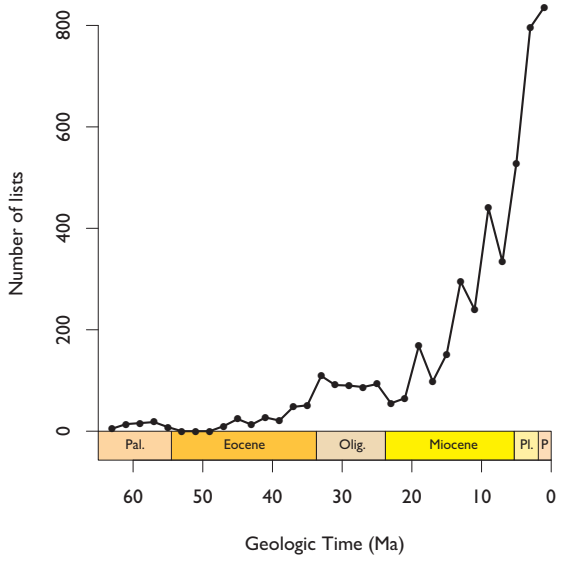


Figure 4.2: Number of lists (of taxa found at a particular depth in a particular borehole) in the *Neptune* database through the Cenozoic Era.

Such temporal sampling biases are common in paleontological datasets, for a

number of reasons. In general, older sediments are less abundant than younger ones. For microfossils from deep-sea drilling cores more specifically, sediments are progressively destroyed as ocean crust becomes subducted by plate tectonic processes, making older sediments less common. Perhaps more importantly, the deep drilling commonly required to reach older sediments is expensive, and requires drilling through younger sediments for which samples are usually also collected. Finally, diatoms undergo a series of diagenetic mineral transitions as burial temperatures and pressures increase, making the preservation of recognizable morphological features less likely with age (DeMaster 2003). In recent decades, paleobiologists have directed much research effort towards developing numerical methods to correct for these sampling biases.

4.2.4 Subsampling methods

The idea at the heart of these methods is that we can obtain a more accurate reconstruction of diversity history if we measure diversity using a standardized, comparable sample size in each time bin. It is important to note that the goal of such subsampling methods is not to get closer to the true absolute values of diversity (as range-through taxon counting does, for example), but rather to reconstruct the relative shape of the diversity curve as accurately as possible. The basic procedure underlying these methods involves randomly drawing items from the full data set of the time bin in question until some quota is reached. The number of taxa in that subsample is then tallied and the process is repeated a large number of times to obtain an average diversity and an associated confidence interval for that time bin. This whole process, in turn, is repeated for all time bins, using the same quota for each.

The well-established subsampling methods all use a uniform quota of items, but differ in how items are drawn and how the quota is set. For example, in classical rarefaction (CR, Miller and Foote 1996), occurrences are drawn from the full dataset until a quota of some fixed number of occurrences is reached. The items drawn can also be taxonomic lists, which in the case of the *Neptune* data

means the list of taxa found in a particular borehole at a particular depth. In this case, the quota can be simply set as a number of lists, in which case the method is referred to as by-lists, unweighted subsampling (UW, Alroy 2000), or as a number of occurrences, (weighted by occurrence subsampling, or OW for short, Alroy 1996). In practice (at least for the *Neptune* diatom data set), these methods give broadly similar results. Subsampling methods were only recently applied to diatoms (Rabosky and Sorhannus 2009), painting a picture of diatom diversification very different from curves generated by RT and SIB tabulations of *Neptune* data (Fig. 4.1). A representative diversity curve generated by OW subsampling (Fig. 4.1 OW), similar to those produced by CR and UW subsampling, still shows a net increase over the Cenozoic Era, but the total increase is only about twofold. Perhaps more importantly, maximum diversity is sited in the late Eocene Epoch, with tabulated diversity falling substantially through the Oligocene Epoch, before recovering through the Neogene Period.

Another method of subsampling, which uses a quota measured by the sum of squared occurrences (O2W), has also been applied to diatoms and results in an even flatter diversity curve with a still more pronounced Eocene-Oligocene diversity peak (Rabosky and Sorhannus 2009). The reasoning that underpins this method (Alroy 2010a) is a little complicated, but can be understood by considering how taxon occurrences are recorded. A paleontologist describing the species in a sample will examine specimens, i.e. individual fossils, recording new species as they are encountered. Because not all taxa are equally abundant, the common taxa will be found quickly, while rare taxa are only likely to be counted after many specimens have been examined. This results in an asymptotic collector curve (or accumulation curve), in which taxa discovered are plotted against specimens examined (or sampling effort expended). The O2W method attempts to account for this non-linear relationship between specimens and species. Unfortunately, diversity estimates under this method have been found to be strongly biased when the geographic structure of diversity changes through time (Bush et al. 2004), so we do not show O2W results here.

A major shortcoming of all of the fixed-quota subsampling methods is that

they can systematically undersample intervals with high diversity, because uniform sampling is not necessarily fair sampling (Alroy 2010a). Consider an ideal case where there has been no change in the true diversity through time, but an increase in sampling results in an apparent increase in diversity: in this situation, fixed-quota methods will theoretically perform well. But if there has been a true increase in diversity in addition to an increase in sampling, these methods may artificially flatten the resulting diversity curve, because more diverse intervals will require more sampling to capture the same proportion of the total (true) diversity. As an extreme example, consider that a population of five species can be sampled completely with 100 occurrences, while the same size sample will underestimate radically the diversity of a population with 500 species. When we allow for the possibility that true diversity can vary widely, it makes intuitive sense to allow the quota of items drawn in subsampling to vary also—and from this perspective, we can also see that the level to which we ought to standardize samples is not a fixed amount of sampling, but some amount of sampling that aims to return a fixed proportion of the total diversity, or coverage.

A recently published subsampling method, shareholder quorum subsampling (SQS), is based on this principle of taxonomic coverage (Alroy 2010a). The method is named by analogy to a corporate shareholder's meeting, where a quorum of shareholders needs to be present such that the sum of their shares meets a threshold proportion of the total shares in the company. In SQS, we can think of taxa as shareholders and their share of the frequency distribution (proportion of total occurrences) as shares. Occurrences are drawn much as before, but rather than stopping at a certain number of occurrences or lists, samples are drawn until a 'shareholder quorum' is reached—that is, until the sum of the frequencies of sampled taxa exceeds some threshold. The number of lists or occurrences it takes to reach this quota is free to vary, making this method philosophically quite distinct from the uniform sampling methods like CR, UW or OW.

In order to ensure that the proportion of frequencies drawn represents a certain proportion of taxonomic coverage, however, we need to know how

taxonomically complete each sample is—that is, we need an estimate of coverage. Another way to think of this problem is to consider the observed frequency distribution to overestimate the frequency of those taxa observed in favor of those taxa not observed, whose frequencies are rounded down to zero. We need an estimate of how much of the underlying, true frequency distribution has been muted by such rounding down. One such estimate, used commonly in ecology, uses the proportion of observations that are singletons, i.e. taxa only seen once in that sample, calculated as Good's *u* (Good 1953). In ecology these observations are individuals, but the approach can be extended to compiled paleontological data where these observations are occurrences (i.e. the presence of a taxon at a particular location and stratigraphic position, irrespective of its abundance; Alroy 2010a). While for the purposes of ecological studies, singletons are those taxa that occur only once in a sample area such as a quadrat, Alroy (2010a) argued that for paleontological data the best analogical equivalent was to count singletons as taxa occurring only in a single publication. We applied a modified version of this basic SQS algorithm to the diatom occurrence data from *Neptune*. Because of the way micropaleontological data are collected (the occurrence of a set of taxa reported over a stratigraphic range), taxa will almost always have more than one occurrence in any given publication. Instead, we used taxa occurring only in a single borehole in place of singletons.

The results for SQ subsampling of the *Neptune* diatom data (Fig. 4.1, SQS) are broadly similar to those of the fixed-quota subsampling methods (Fig. 4.1, OW), suggesting that true diversity increased only slightly over the Cenozoic Era. Much as under the fixed-quota methods, peak diversity is reached in the latest Eocene/earliest Oligocene, but under SQS this peak is exaggerated, suggesting that diversity then was significantly higher than today—similar to the results of the O2W method (Rabosky and Sorhannus 2009).

In spite of the obvious sampling bias in the *Neptune* data, the largely stationary view of Cenozoic diatom diversity suggested by subsampling methods has not been universally accepted by micropaleontologists. A potential vulnerability of subsampling methods, including SQS, is that they may not give accurate results if

there are large changes in relative abundance structure (or evenness) through time (Lazarus et al. 2012a). We can understand this problem by considering the relative frequency distribution of a time interval—a rank-ordered plot of the proportion of occurrences of each taxon (such that the extent along the x-axis represents total diversity). Fixed-quota subsampling methods can be thought of as sampling all taxa falling above a threshold 'veil line' of some relative frequency (Alroy 2010a). The failing of these methods, addressed by SQS, can be visualized by considering what happens if the diversity increases, but the shape of this distribution stays the same: because each taxon now has a smaller relative frequency, a greater proportion of the taxa falls under the veil line, underestimating diversity under subsampling. Under SQS, a constant area under this curve is sampled, so even if the total diversity increases, the same proportion of the frequency distribution will be recovered—and, if the shape of the distributions stays the same, the same proportion of total diversity. If the shape of the frequency distribution were to change drastically, however, SQS might not work as well.

Empirically, the diatom occurrence data in *Neptune* do show a change in frequency distribution from more even to more uneven, and it has been argued that these changes may cause subsampling methods (including SQS) to mask a true rise in diversity (Lazarus et al. 2012a). If we imagine SQS subsampling to recover a fixed area under a rank-ordered relative frequency distribution (see supplement to Alroy 2010c), the area under a flat curve (an equitable frequency distribution) will sample a greater proportion of the total diversity than the same area under a hollow curve (an uneven frequency distribution). Lazarus et al. (2012a) apply an empirical correction factor to account for the changes in frequency distribution and recover a rise in diatom diversity more similar to the canonical view. A similar correction factor with even greater leverage is used to account for an increase in provinciality through time, particularly regarding the development on an endemic polar fauna (Lazarus et al. 2012a).

Lazarus et al. (2012a) marshal further support for the conventional view of diatom diversification from a catalogue of about 500 diatom species' ranges

compiled from both marine and land-based sections under expert curation against taxonomic and stratigraphic error. The curve generated is similar in form to the canonical diatom diversity curve (Spencer-Cervato 1999), albeit showing a net increase that is slightly less steep. This data set has certainly been better flushed of "RATs" (the sorts of errors described in the section on data quality above) than *Neptune*, but the question of sampling bias arguably remains: while there is no strong correlation in this compilation between diversity in a time bin and the number of publications from which this diversity is derived (Lazarus et al. 2012a), the relationship between a taxonomic or biostratigraphic publication and the amount of sampling it represents is not clear and not necessarily fixed.

To summarize, the taxonomic window on diatom diversification provides an uncertain picture of Cenozoic diatom evolution. Interpreted at face value, the fossil record suggests a steep Cenozoic rise in species richness, whether from deep-sea occurrences in the *Neptune* database (Spencer-Cervato 1999) or from a biostratigraphic catalogue of first and last appearances (Lazarus et al. 2012a). When the stark secular rise in the amount of available data is taken into account using item quota (Rabosky and Sorhannus 2009) or SQS subsampling methods, however, a more stationary pattern emerges, showing at most a modest overall increase in species richness and peak diversity around the Eocene/Oligocene boundary. With changes in relative abundance potentially biasing the results of these subsampling methods, we are left with a level of uncertainty about the true diversification history of the diatoms. Recalling that there is another window on diversification, however, we turn to the history of diatom morphological disparity to gain another perspective on this question.

4.3 Reconstructing Evolution in Shape Space

In common paleobiological usage, disparity describes a quantification of morphological differences among organisms (Wills 2001, p. 56). Unlike species richness for diversity, there is no singular metric for disparity; commonly used measures can be more easily understood in the conceptual framework of

morphospace—a mathematical construct used to quantify and describe organismal morphology.

4.3.1 Morphospaces

Morphospaces are n-dimensional mathematical spaces describing the form of a group of organisms. As such, morphospaces are an example of what, in the context of ecology, Lewontin (1969, p. 13) called "the concept of the vector field in n-dimensional space," which he described as "the most fundamental [concept] we have for dealing with the transformations of complicated dynamical systems in time." Familiar, conceptually related notions include adaptive landscapes (Wright 1932) and niche space (Hutchinson 1978, p. 158), but rather than gene alleles or ecological variables, the axes of morphospaces represent morphological characters or parameters. Each point in morphospace represents a particular, unique morphology, and it can either be occupied (i.e. represent a morphology actually realized by an organism) or not.

With this framework in mind, we can consider the morphological disparity of a group as a description of how the group is distributed in morphospace—are the taxa spread out widely (signifying large morphological differences) or clustered together (signifying morphological similarity)? As discussed in more detail below, this spatial distribution of taxa can be quantified in a number of ways, leading to multiple metrics of disparity. Before considering how to measure morphospace occupation, however, it is worth briefly examining the different ways in which morphospaces can be constructed.

Morphospaces are often divided into two kinds, those whose axes are parameters of a shape-generating function, called generative or theoretical morphospaces, and those whose axes are measurements of organisms, called empirical morphospaces (McGhee 1999). Theoretical morphospaces generally have only a few axes and thus a small number of dimensions that is easy to visualize; the first and best-known example is the Raup and Michelson (1965) morphospace of coiled shells. Empirical morphospaces, in contrast, often have a

very large number of axes (representing a large number of measured morphological characters) and generally require an ordination procedure such as principal components analysis (PCA) or principal coordinates analysis (PCO) for visualization and analysis, an approach pioneered by Foote (1989). Because of this, empirical morphospaces have been described as having axes that are data-dependent or unstable, since different measurements of the same morphology will result in different ordinated axes (McGhee 1999). Seen from a more general perspective, however, the distinction between theoretical and empirical morphospaces can become conceptually and mathematically blurred if the latter are considered in their full, unordinated dimensionality (sometimes called "raw morphospaces," Eble 2000b): the number of axes could then be seen as the most significant difference between the two. From this perspective, both sorts of morphospace can be used to investigate the realms of unrealized as well as realized morphologies—although theoretical morphospaces can undoubtedly generate a wider range of unrealized form than empirical morphospaces can.

4.3.2 Limitations of theoretical morphospaces

While there is broad consensus that theoretical morphospaces are preferable because their use of explicit, measurement-independent growth models that allow one to explore a wider range of unexplored as well as impermissible forms (e.g. Erwin 2007), their application is unfortunately not always possible (McGhee 1999, p. 26). Growth models for theoretical morphospaces are more readily devised for organisms with accretionary or branching growth (e.g. Raup and Michelson 1965; Niklas 1999), but mathematical shape models with a reasonable number of parameters can only reproduce so many aspects of form. The applicability of generative morphospaces with a small number of parameters is thus limited in a two ways that are well illustrated by the case of the diatoms: the range of overall forms that can be generated, and the difficulty of including complex and higher-order morphological features.

4.3.3 Previous diatom morphospaces

The diversity of fundamental forms that can be generated by a mathematical model with a few parameters is limited. In diatoms, for example, capturing the great variety of different symmetries of the valve in plan view alone (circular-elliptical, triangular, rectangular, curved, isopolar or heteropolar, and so on) in a generative model would require many parameters, and even then the plan-view outline shape says nothing about the obviously important three-dimensional shape of the valve. Generative shape models that have been developed for the diatoms are thus by necessity limited both in terms of covering only a subset of the full taxonomic and morphological diversity, and in terms of describing a subset of the overall frustule morphology. Examples include models for a particular species (Stoermer and Ladewski 1982) or genus (Mou and Stoermer 2004), a more widely applicable model describing only valve outlines (Arita and Ohtsuka 2004), and a model based on 3D parametric equations limited to a group of asymmetrical pennate diatoms (Pappas 2005). While generative morphospaces of this nature have been profitably applied to questions of taxonomic distinction or morphological evolution within particular groups, they capture neither the total diversity of overall diatom form, nor the higher-order features of diatom morphology such as pore arrangement, spines, processes, or the raphe—even though these may well be of biological and evolutionary significance.

We are thus led to an empirical morphospace approach in trying to understand the Cenozoic evolution in morphospace of the marine diatoms in as a whole. This approach has been successfully applied to many other groups with complex morphologies, highlighting important features that are hard to represent with simple geometric models (e.g. Foote 1995a; Lupia 1999; Boyce and Knoll 2002). Under the auspices of the PlanktonTech initiative, we conducted an empirical morphospace study of Cenozoic marine planktonic diatoms, the detailed methodology and results of which are published elsewhere (Chapters 2 and 3). Summarized briefly, this approach involved quantifying the morphology of

Cenozoic diatom taxa using a large number of morphological characters to construct a morphospace, applying an ordination procedure to visualize this morphospace, and then populating the morphospace through time based on the fossil record. Since the *Neptune* database is a readily available compilation of fossil diatom occurrences, we used it to populate the morphospace. Because an analysis at the species level would be intractable (there are over 1,000 diatom species in *Neptune*), we chose to work at the genus level; our final analysis included 140 genera.

4.3.4 Choice and coding of characters

We used descriptions of frustule morphology and taxonomic descriptions of the genera in *Neptune* to compile a list of morphological characters. In formulating these characters, we were careful to strictly describe morphology independent of taxonomy or phylogeny—meaning that structures were quantified by their similarity in form regardless of whether they are equivalent in development or evolutionary origin. In some cases, this meant bridging substantial gaps in the nomenclature used to describe structures in different groups within the diatoms, for example the terms applied to the arrangement of pores in centric diatoms ("areolation") versus in pennate diatoms ("striation"). This strictly morphological approach distinguishes morphospace analysis from morphometric approaches in which the importance of choosing homologous characters is often strongly emphasized (e.g. Rohlf and Bookstein 1990). It has the advantage of allowing the evolutionary exploration of form on its own merit, independent of how this form is achieved phylogenetically or developmentally. Our final analysis included 100 morphological characters, coded as unordered, discrete character states (including many binary characters), allowing us to account for both intrageneric variation and the categorical nature of many characters (such as the presence or absence of a raphe, a slit along the valve face of some diatoms that enables locomotion).

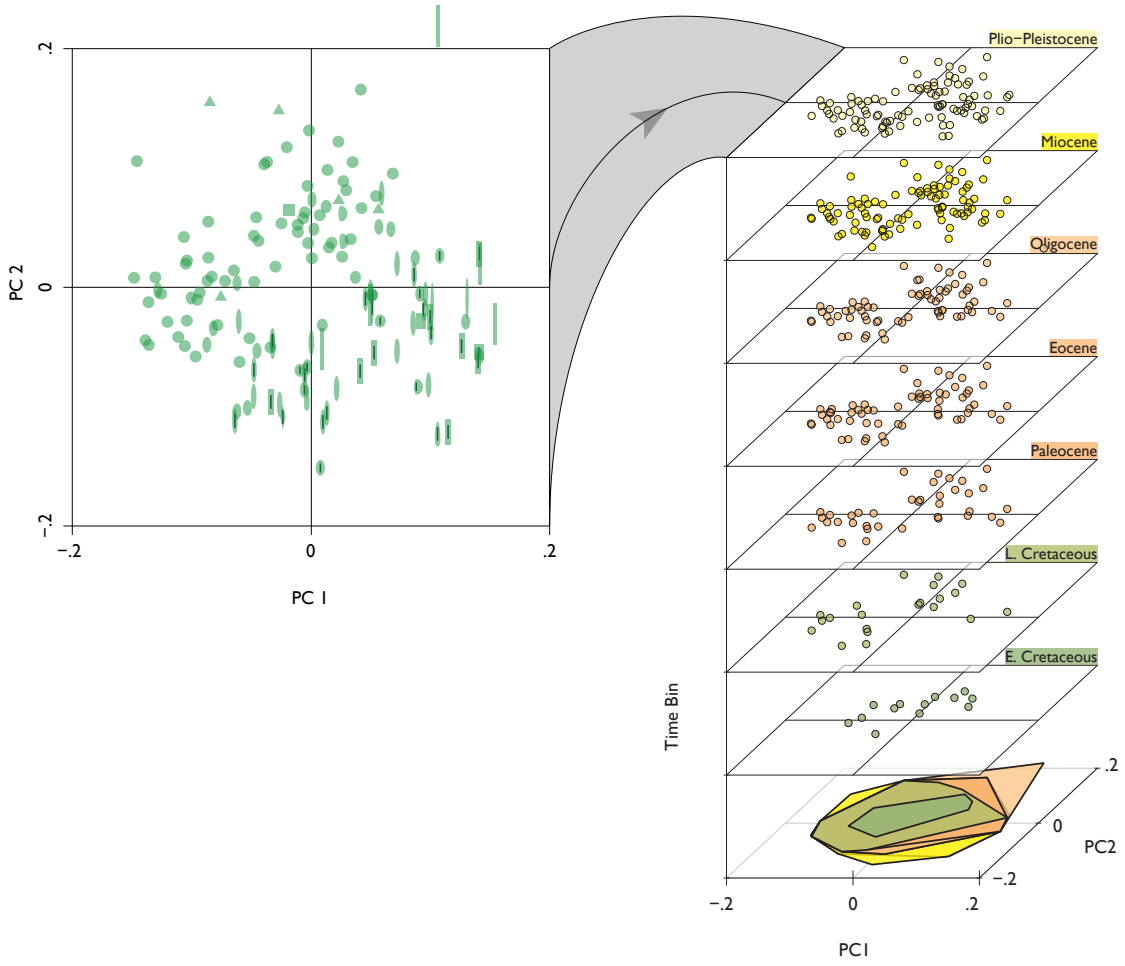


Figure 4.3: Left: first two principal coordinate (PCO) axes of marine planktonic diatom morphospace; plot symbols reflect states morphological characters describing valve shape in plan view and presence/absence of raphe. Right: the same plot populated through time using occurrence data from the *Neptune* database; colored polygons below show convex hull areas enclosing points in time bins above of the corresponding color.

4.3.5 Ordination

The resulting morphological character codings for each genus define a 100-dimensional, categorical morphospace, a "raw morphospace" in the sense alluded to above. Visualizing this morphospace requires that we find a lower-dimensional projection of the relative locations of our genera, something that can be accomplished with PCO, a method analogous to the more familiar PCA that works with unordered discrete characters. This transformation results in a representation of the morphospace along continuous axes, the first two of which (the two capturing the greatest amount of the information in the full-dimensional morphospace) are shown on the left in Figure 4.3.

4.3.6 Visualization

Empirical morphospace plots resulting from ordination procedures like this one can be hard to interpret if the only information presented is the location of genera on ordinated axes of mysterious meaning. Providing representative images of morphologies at selected locations is helpful, but because we have morphological data for each of the points, it can make sense to modify the shape of plot points themselves to reflect morphological character states. In the left panel of Figure 4.3, we have represented the states of three characters relating to the plan (valve) view shape of each genus: the overall shape (elliptical, triangular, square, or ovate), the aspect ratio, and the presence or absence of a raphe. The plot shows equant forms toward the top left, and elongate forms with and without raphes toward the bottom right. This division reflects the largest-scale taxonomic division within the diatoms—centrics versus pennates—and raises the question to what extent phylogenetic structure is evident in morphospace.

4.3.7 Morphospace and phylogeny

Comparing the proximity of genera in morphospace to their proximity on a molecular phylogenetic tree shows only a very weak correlation; phylogenetic proximity at finer levels of resolution beyond the centric-pennate divide is not a

good predictor of proximity in morphospace (see Chapter 2). For example, the two major subdivisions within the pennate diatoms (raphids and araphids) do not appear to occupy distinct regions of morphospace. This may seem surprising, considering diatom phylogenies from before the molecular era broadly agree with molecular ones. It helps to remember, however, that morphological phylogenies identify key features with defined polarities, called synapomorphies, signifying inclusion in groups; this morphospace, in contrast, consists of equally weighted characters.

These results seem to suggest that the major groups of diatoms iteratively recolonized already-occupied regions of morphospace. In the case of the raphe, for example, this might make sense if we consider that—in providing for locomotion—it represents a key innovation in the radiation of diatoms in benthic and terrestrial environments. The taxa in our analysis come from the marine plankton, however, where—in the absence of substrates upon which to locomote—this innovation may have been of relatively little consequence. As a result, raphid diatoms in the plankton may have radiated to fill ecological niches indistinct from those occupied by their araphid cousins, resulting in overlapping occupation of morphospace.

Molecular clocks suggest that the four major groups of diatoms (the raphid and araphid pennates and two groups of centric diatoms, the radial centrics and bi- and multipolar centrics) had diverged by the late Cretaceous Period (e.g. Kooistra et al. 2007). Given the apparent overlap of these groups in morphospace, we might expect to see broad morphological stasis across the Cenozoic Era. Fortunately, we can further explore this question by extending our morphospace back through time using the fossil data from *Neptune*.

4.3.8 Morphospace through time

The right hand side of Figure 4.3 shows the occupation of the two-dimensional morphospace plot on the left through geologic time, tilted to an oblique view, for several Cenozoic time bins (with the youngest at the top). The polygons at the

bottom of the plot are convex hulls enclosing the points in each of the time bins above in the corresponding colors; convex hulls being the shape that would be made by stretching a rubber band around the points in a time bin and thus describe the extent of occupied morphospace. While the Cenozoic assemblages cover a larger area of the plot than the Cretaceous ones, the area does not increase very much through the Cenozoic Era. One interesting observation is that the quadrant at the center of pennate diatom morphology (the bottom right) is sparsely occupied up until the Miocene. This general impression of how morphospace has been occupied through time can be quantified much more rigorously by calculating metrics of disparity for each time bin in the *Neptune* data set.

4.3.9 Metrics of disparity

Many different ways of measuring disparity have been devised. A thorough review of these different metrics is beyond the scope of this chapter (for more detailed treatments see Wills et al. 1994; Ciampaglio et al. 2001; Erwin 2007), but a key point is that these metrics do not all describe the same aspects of morphospace occupation. To illustrate this notion, we present two disparity metrics here: convex hull volume and mean pairwise distance. The former is a higher-dimensional extension of the rubber band method introduced in Figure 4.3; instead of measuring the area enclosed by the polygon, the volume or hypervolume enclosed in three or more dimensions is calculated. While convex hull volume is a measure of the total extent of morphospace occupied, the latter metric—as implied by its name—measures the average distance between pairs of taxa, measured as the proportion of character state mismatches out of possible matches.

It is not hard to see that because these metrics describe different aspects of morphospace occupation—the total extent of occupied morphospace and the degree of dispersion of taxa in morphospace, respectively—they can give different answers regarding disparity trends through time. In the case of diatom

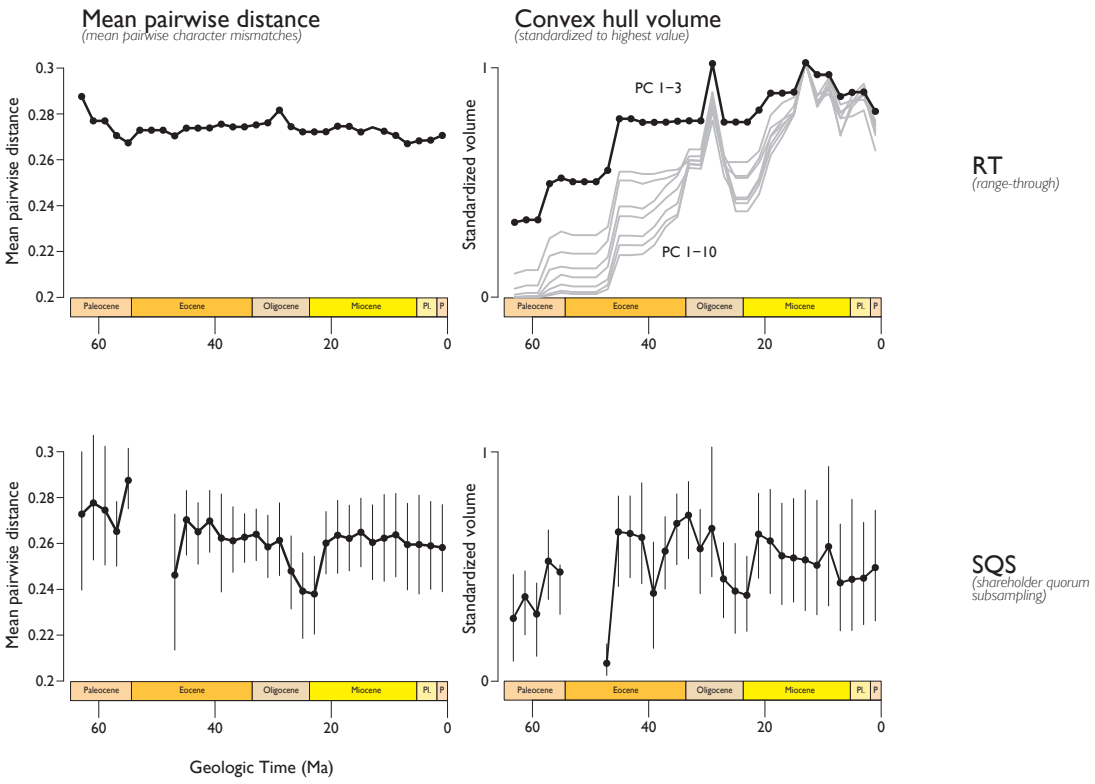


Figure 4.4: Metrics of diatom disparity through time under in-bin sampling of the *Neptune* occurrence data taken at face value (top row) and shareholder quorum subsampling of the same dataset (bottom row). Plots on the left show the average dissimilarity (or morphological distance) between taxa measured as the proportion of character state mismatches to possible matches, plots on the right show the volume of convex hulls enclosing genera present in each time bin (in the top plot, 3–10 dimensions, in the bottom plot, 3 dimensions only).

morphospace, for example, mean pairwise distance (Fig. 4.4, top left panel) stays roughly constant through time, while convex hull volume (Fig. 4.4, top right) shows a substantial increase (regardless of how may dimensions are considered in the volume calculation). While these results would appear to be at odds if we were to consider disparity as a monolithic concept, inspection of the stacked morphospace plot in Figure 4.3 makes clear which aspects of morphospace occupation they describe: on the one hand, the total extent of occupied morphospace is increasing as new genera expand into previously unoccupied morphologies, while on the other hand, the number of genera is increasing, leading to a roughly constant (or slightly increasing) "packing" of genera into the occupied region. We can thus think of mean pairwise distance as representing something of a density measure and convex hull volume as the total extent of occupied morphospace. As is the case in Figure 4.4, the density (mean pairwise distance) can remain roughly constant as occupied morphospace (convex hull volume) expands, but only if there is a commensurate rise in diversity.

4.3.10 Sampling

This increase in the number of genera in morphospace raises an interesting question about sampling. Since we spent much of the first half of this chapter voicing concern about the possibility that the evident rise in taxonomic diversity may simply be an artifact of sampling, it seems natural to ask: could sampling also affect metrics of morphological disparity? Although we introduced disparity as a different window onto the diversification history of a clade, it is based on the same fossil record and is thus subject to the same geological sources of bias. Studies on the effect of sampling on disparity metrics have suggested the use of mean pairwise distance, which has been observed to be more robust to variations in sampling than other disparity measures (Foote 1992; Ciampaglio et al. 2001; Butler et al. 2012). As we have just seen, however, considering one disparity metric alone limits analysis to just one aspect of morphospace occupation. While it is always possible to consider the effect of increasing numbers of taxa in a

morphospace indirectly, for example, by rarefaction to a standardized number of taxa (Foote 1992), we can do better in the unique situation where a morphospace has been populated using a database of fossil occurrences. In this case, we can apply those subsampling metrics (discussed above) developed for addressing sampling bias in studies of taxonomic diversity to our morphospace, and calculate metrics of disparity under subsampling.

The results of applying SQS to the diatom morphospace and calculating our two chosen disparity metrics are shown in the bottom half of Figure 4.4. Mean pairwise distance does not appear to be significantly impacted by the subsampling exercise, lending further support to the notion that this metric is robust to sampling variations. For convex hull volume, in contrast, the increasing trend over the Cenozoic seems to all but disappear, revealing a much more stationary pattern through time. When sampling is taken into account, then, disparity metrics paint a picture of relative Cenozoic stasis in diatom morphological evolution that stands in agreement with both the comparison between morphospace and phylogeny and the results of subsampled taxonomic diversity.

4.4 Synthesis

The fossil record provides us with two windows onto the Cenozoic diversification history of diatoms, through taxonomic diversity and morphological disparity. When read at face value, the record suggests steep Cenozoic diversification from both a view through diversity and through convex hull volume, a disparity measure describing the total extent of occupied morphospace; mean pairwise distance, in contrast, suggests a stationary pattern through Cenozoic time. When secular trends in sampling intensity are taken into account using both well-established and new subsampling methods, however, the records through both windows broadly suggest stasis, a pattern also predicted from a comparison of morphospace and molecular phylogeny. While disparity as measured by mean pairwise distance seems to be robust to sampling, the other

results—disparity as well as diversity—hinge upon whether we believe that subsampling algorithms do a better job at uncovering true diversity history than a literal reading of the fossil record, or whether they simply trade sampling bias for another bias resulting from changing relative abundance distributions.

A similar pattern has been discovered in another group of open ocean microfossils, the planktonic foraminifera: a literal reading of their deep-sea record also suggests a steep and roughly monotonic rise in species richness through Cretaceous and Cenozoic time, but, when sampling is accounted for by either subsampling or modeling, a much more gradual rise is recovered, with peak diversity in the Cretaceous (Lloyd et al. 2012a). And the fossil record of cocclithophorids also contrasts a literal reading of the Cretaceous-Cenozoic record with results obtained when sampling is taken into account (Lloyd et al. 2012b).

Without the compelling empirical suggestion of a Cenozoic rise in diatom diversity, it is worth considering whether an unfettered diversification would be expected a priori. The answer is not immediately obvious, but surely requires consideration of the relationship between phytoplankton diversity and both physical and chemical oceanography—as highlighted, for example, by global marine ecosystem models that implicate the role of resource availability and dispersal in controlling phytoplankton diversity (Follows et al. 2007). While these oceanographic factors are undoubtedly linked to climate, how exactly changes in climate would be expected to affect diatom diversity is a question deserving of further attention.

If the pattern of relative Cenozoic stasis in diversity and disparity to which the results presented here point is accurate, most of the marine planktonic diatoms' diversification was a Mesozoic to earliest Cenozoic event, perhaps prompting a reexamination of evolutionary narratives in which a Cenozoic rise in diatom diversity features prominently. In many of these narratives, however, diatom diversity merely stands as a proxy for diatom participation in the silica cycle, yet the number of taxa is only one factor in their importance to silica cycling. It is also conceivable that there were changes through time in diatom abundance or

the rate of diatomaceous sediment deposition. Such changes could, at least theoretically, be independent of diversity; consider, for example, that in the Southern Ocean diatom ooze belt, perhaps the most important area of diatomaceous silica deposition today, sediments are dominated by just one species, *Fragilariopsis kerguelensis*, constituting up to 60 to 90% of total diatom abundance (Zielinski and Gersonde 1997). In addition to its abundance, *F. kerguelensis* is rather heavily silicified, illustrating also the potential role of changes in silicification to the diatoms' biogeochemical impact. The origination of even a small number of such numerically dominant or robustly silicified taxa could potentially expand the diatoms' role in the silica cycle to an extent much greater than the concomitant taxonomic diversification. Indeed, as described in the discussion of subsampling methods above, frequency distributions in the *Neptune* database indicate an increase in dominance through time compatible with such a scenario.

From the perspective of diatoms serving as starting points for biologically inspired design, one implication of an early exploration of morphospace might be that fossil morphologies ought to be considered alongside those of extant diatoms. Particularly if the maximum range of diatom form was achieved early in the Cenozoic, fossils from that time period may provide a range of biological constructions of engineering value not available in recent forms.

There are also important limitations of this study that must be considered. Constructing a morphospace capable of representing the full taxonomic and temporal sweep of a clade as large and diverse as the diatoms requires trade-offs in the level of morphological detail that can be recorded. For example, changes in the degree of silicification of diatom frustules are not well captured by the morphological characters in this study, since these are not necessarily visible in those characters that can be coded cohesively at the genus level. If predictions from the fossil record of radiolarians (Lazarus et al. 2009) and semi-quantitative observations of the diatom fossil record (Finkel and Kotrc 2010) hold true, diatoms ought to show a reduction in silicification over the Cenozoic Era, a pattern of interest to engineers seeking structures that maximize strength with

minimal use of constructional material. Such trends may be best investigated by looking at morphological changes within long-ranging genera (such as *Stephanopyxis)*, where insight might be gained to how nature does more with less.

Eine Messerspitze voll von dem feinen kreideähnlichen Radiolarien-Schlamm, der Tausende von Quadratmeilen des Ocean-Bodens bedeckt, enthält gewöhnlich mehrere Hundert verschiedene Arten, und Tausende von Individuen. Die sorgfältige Untersuchung dieser wundervollen Schätze—eine „mikroskopische Gemüths- und Augen-Ergötzung" ersten Ranges—hat mich über ein Decennium hindurch gefesselt.

Ernst Haeckel

5

Changes in Silicification within Cenozoic Radiolarian Lineages

Abstract

The Cenozoic fossil record of radiolarians shows an exceptionally clear macroevolutionary decline in silicification, reflecting changes in test thickness and porosity at the assemblage level. The evolutionary mechanisms underlying this pattern have remained unclear. I addressed this question by examining changes in silicification along three well-documented evolutionary (anagenetic) radiolarian lineages: Stichocorys, Didymocyrtis, *and* Centrobotrys. *Using samples from four tropical Pacific DSDP drill sites, I made 6790 measurements of 44 morphological parameters under transmitted light microscopy, stored in a custom relational database. I used geometric models of test morphology based on cones, spheres, cylinders, and portions*

thereof in order to calculate percent silicification from these measurements. The resulting trends were fit to three evolutionary models—representing directional change, random walk, and stasis—by maximum likelihood. Two lineages support the random walk model, while one supports the model for stasis, suggesting that macroevolutionary processes above the species level are responsible for assemblage-wide decrease in silicification. The differences among lineages can be explained by biological differences in the role of the test in feeding ecology. No relationship was found between pore area and thickness, unlike in diatoms, suggesting that the radiolarian test plays a different biological role than the diatom frustule. Although the results point toward selection among, and not within, lineages, three caveats caution against ruling out a role for anagenesis: (1) two of the lineages do show net silicification changes that could add up to an overall decrease in silicification over many lineages; (2) variations in species abundance may play a role in explaining the whole-assemblage pattern; and (3) most of the change in the assemblage-level trajectory occurs in the Paleogene, so we might not expect to see directional change in the Neogene lineages examined here. A turnover event at the Eocene/Oligocene boundary could explain the assemblage-level pattern if there was biased extinction of highly silicified lineages and biased origination of lightly silicified lineages.

5.1 Introduction

The Cenozoic coevolution of radiolarians and diatoms describes a narrative that exemplifies the emerging discipline of geobiology: global-scale changes resulting from the interactions among the biological participants in the geochemical silica cycle. Both radiolarians and diatoms make preservable hard parts out of silica and are thus tied by physiological requirement to the silica cycle. The key observations from the Cenozoic linking their evolutionary histories were first assembled by Harper and Knoll (1975), who showed that a documented Cenozoic decrease in radiolarian test weight (Moore 1969) paralleled diatom diversification, suggesting that radiolarian evolution reflects the rise of diatoms to their present-day dominance of the marine silica cycle.

The changes in test thickness and porosity postulated to be responsible for this change in test weight were later documented to indeed change over the course of the Cenozoic Era (Lazarus et al. 2009), determining that the trends in body size observed in related plankton groups (e.g. Schmidt et al. 2004) are not found in radiolarians, and are not thus responsible for changes in test weight. Furthermore, these changes were observed in low latitudes but not in the Southern Ocean, where the dynamics of the modern-day silica cycle keep concentrations of dissolved silica high, further supporting the conclusion that radiolarian evolution was driven by changes in silica availability effected by diatom evolution.

The documentation of such an exceptionally clear macroevolutionary shift in radiolarian silicification—a near monotonic decline from above 16% to around 6% of test volume (Lazarus et al. 2009, Fig. 2)—raises the question of what evolutionary mechanisms underlie this pattern. Did change in silicification occur along anagenetic lineages? Did more heavily silicified lineages go extinct while more lightly silicified lineages persisted or originated? Was the assemblage-level pattern a result of either process exclusively, or did both occur? Answering these questions comprehensively is beyond the scope of a single thesis chapter, as it would require both substantial alpha taxonomy and establishment of evolutionary relationships, both of which remain unclear for a majority of radiolarian forms at the species level (Lazarus 2005, and pers. comm.). We can, however, answer the first of these questions by examining changes in silicification along individual evolutionary lineages. In this chapter, I examine silicification change in three relatively well-documented evolutionary (anagenetic) lineages of Cenozoic radiolarians.

Seeking to understand the mechanisms behind macroevolutionary trends is of broad paleobiological relevance to the long-standing debate about how the hierarchical levels of evolution interact (Jablonski 2007). Much of this debate has focused on the question of whether evolution above and below the species level differ fundamentally in pattern and process. As Doug Erwin has pointed out, however, "as is so often the case in evolution, the interesting question is not, is macroevolution distinct from microevolution, but the relative frequency and

impact of processes at the various levels of this hierarchy" (Erwin 2000). By examining the species-level mechanisms underlying the clade-level, macroevolutionary pattern observed in the Radiolaria, I aim to contribute one well-documented example to our understanding of how evolution works across hierarchical levels.

5.2 Background

5.2.1 Mechanisms Underlying Macroevolutionary Patterns

For the last three decades of the 20^{th} century, the debate over punctuated equilibria (sparked by Eldredge and Gould 1972) focused much paleobiological research effort on documenting the temporal distribution of morphological change in the fossil record. The punctuated equilibria model proposes that morphological change is concentrated in relatively short periods of time associated with speciation and interspersed with relatively long periods of stasis, giving rise to the now-famous motto: "stasis is data" (borne out, in a different context, in the previous chapters of this dissertation). In spite of the extensive research triggered by this debate, however, there remain relatively few published examples of within-lineage changes in the context of well-documented macroevolutionary trends.

In characterizing species as largely stationary throughout their stratigraphic ranges, the punctuated equilibria model posits that evolutionary trends originate above the species level, describing "macroevolution as the differential success of certain species (and their descendants) within clades" (Gould and Eldredge 1993). This differential success can arise in a number of ways and has been documented in a variety of studies. For instance, it can result from a difference in speciation rates (e.g. between gastropods with planktotrophic and nonplanktotrophic larvae, Hansen 1982), a difference in extinction rates (e.g. between keeled and unkeeled foraminifera, Norris 1991), or an asymmetric increase in variance from a bounded starting point (like the passive increase in

rodent size from a small origin shown by Stanley 1973 or examples reviewed by Gould 1988).

Prior to the recognition of these mechanisms above the species level, the default explanation for macroevolutionary trends was the aggregation of gradual morphologic change within species by anagenesis. Gould summarized the broadened perspective like this: "[I]f we view species as stable entities for most of their geological existence, not as temporary names for transient states in the great and continuous flux of life, then we must interpret trends differently" (Gould 1988). In some cases, however, species are precisely "temporary names for transient states." For example, the Cenozoic radiolarian species defined in the second half of the 20^{th} century by careful stratigraphic tracing of lineages are, in many cases, more or less arbitrarily drawn boundaries on gradually transforming morphologies chosen for their biostratigraphic utility (see the section on "morphotypic" and "evolutionary" limits of species in Riedel and Sanfilippo 1971, p. 1530, for example, or Lazarus 2005). Indeed, species defined in this way are sufficiently common across the paleontological literature to deserve their own term, "chronospecies" (e.g. Stanley 1978). Such anagenetic changes between species have been documented in great detail (e.g. the radiolarian lineages *Buccinosphaera invaginata*, Knoll and Johnson 1975 and *Pterocanium prismatum*, Lazarus 1986).

In some cases, within-lineage changes have been documented in the context of macroevolutionary studies. Many of the best-known macroevolutionary trends involve body size, having attracted the interest of paleontologists since at least the 19^{th} century, when the perceived tendency for body size to increase through a clade's history was named "Cope's rule" (after E.D. Cope). This phenomenon was studied, for example, in a census of body size in Cretaceous mollusks at the genus level (Jablonski 1997). In that study, the maximum size of species within each genus was tracked at the beginning and end of the study interval and directional size increase was found to be no more common than either directional size decrease or increase in variance. Size changes in ancestor-descendant pairs of taxa have also been studied in mammals (Alroy 1998) and planktonic

foraminifera (Arnold et al. 1995). In a broad survey of Jurassic bivalves, Hallam (1978) found evidence of phyletic size increase in *Gryphaea* (better known as "Devil's toenails"), not just between successive species in the lineage, but within species as well.

Probably the most iconic macroevolutionary size trend—and perhaps the most widely discussed example of macroevolution—is the Cenozoic increase in body size among horses. The simplistic view of a steady, linear march of consecutively larger species painted by museum exhibits of the last century has long been rejected (see review by Gould 1992, for example). Prothero and Shubin (1989) ruled out orthogenesis in Oligocene horses, instead finding size increases to result from a series of cladogenetic events, with the species involved showing morphological stasis over millions of years. Other studies of fossil horses, however, have reported chronospecies with gradually evolving sizes (Gingerich 1989) and phyletic changes in size within anagenetic lineages (MacFadden 1985; also see review in MacFadden 1994, p. 170).

Over the course of much of the punctuated equilibria debate, the identification of stasis or directional evolution in paleontological data sets was made qualitatively. Only relatively recently have time series of morphological measurements been tested statistically to determine their correspondence to explicit mathematical models of evolution. The most sophisticated statistical approach to date was introduced by Hunt (2006), who used maximum likelihood methods to fit explicit mathematical models of stasis, directional evolution, and Brownian motion (a random walk) to records of evolutionary change. In an analysis of 250 sequences of evolving traits, Hunt (2007) found that only 5% were fit best by the model of directional evolution; the remaining sequences were split roughly evenly between stasis and random walks (though directional change appeared to be more common in planktonic than benthonic organisms). However, the sequences compiled in the study are considered in isolation and not in the context of observed macroevolutionary trends. In this study, I apply Hunt's method to sequences of morphological measurements collected specifically within the context of an observed macroevolutionary pattern, that of

declining radiolarian silicification over the course of the Cenozoic Era.

5.2.2 Radiolarian Silicification

Relatively little empirical work has been published measuring the degree of radiolarian silicification. In the first study of radiolarian silica use through the Cenozoic era, Moore (1969) weighed prepared radiolarian slides and divided this by the number of individual specimens on the slide to obtain an average test weight for each assemblage. Forty years later, Lazarus et al. (2009) took a different approach, measuring thickness, porosity, length, and width of radiolarian tests and calculating the volume of silica used with simple geometric models (perforate cones and spheres). This approach allowed for the disentanglement of test size and silicification, both of which can contribute to test weight, and showed that the weight trend was driven by silicification, not size.

A similar approach, using simple geometric models to calculate the volume of test silica from measurements of test morphology, has since been applied to Holocene sediments to determine which species are responsible for vertical opal flux in the Southern Ocean (Jacot Des Combes and Abelmann 2009), albeit using values from single specimens to represent entire species, and visually estimating rather than measuring porosity. Since the aim of that study was process-oriented, no secular data were reported. While the study showed that the volume of silica can vary among species by an order of magnitude, these results did not correct for variation in size to isolate the degree of silicification.

The results of the study by Lazarus et al. (2009) show the size-independent degree of radiolarian silicification through time, but only at the whole-assemblage level for all of the Radiolaria in a given sample. The data collected in the course of that study, however, do allow the results to be broken down into families. The results for four common families are plotted in Figure 5.1. These results show that the trend in declining silicification over the Cenozoic Era observed in the whole assemblage (Lazarus et al. 2009) is also seen in individual families.

Two important factors preclude interpreting the family-level trends as

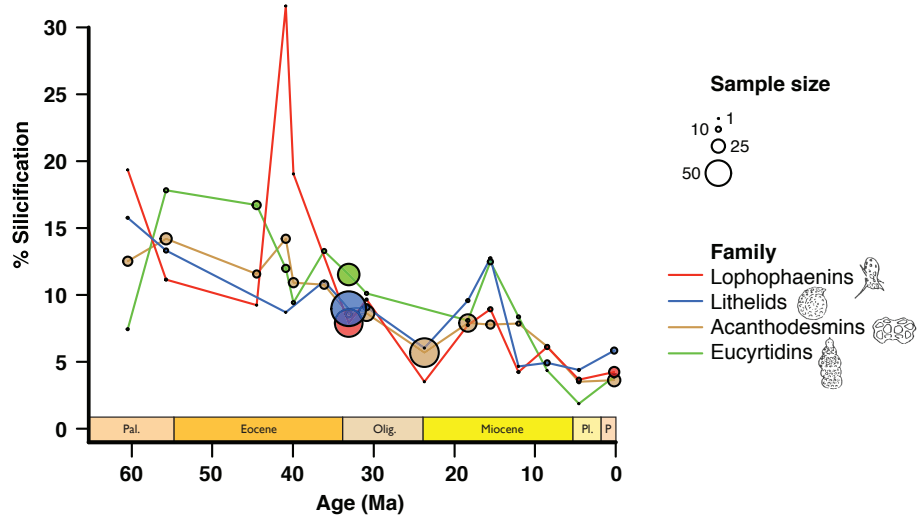


Figure 5.1: The degree of silicification measured in the study by Lazarus et al. (2009) broken down by family, for four more common families. While the results for these families show a similar pattern to the whole-assemblage level, families are likely non-monophyletic, and sample sizes are small.

indicative of evolutionary patterns within clades constituting the Radiolaria. Firstly, sample sizes for individual families are small, because measurements for the study were taken with assemblage-level statistical significance in mind; also, the geometric models used in calculating silicification are crude because they were chosen to be widely applicable to many morphotypes. Secondly, and more importantly, the bulk of radiolarian taxonomy is non-cladistic, particularly at suprageneric levels. The vast majority of radiolarian taxonomy at all levels is based on the monographic work by Haeckel (1887), using almost exclusively characters of test geometry, symmetry, and shape. Subsequent revision of this taxonomy has been hampered by Haeckel's failure to designate type specimens; however, the taxonomic work that has been carried out since suggests that many, if not most, of Haeckel's taxa are para- or polyphyletic (reviewed in Lazarus 2005).

In light of these shortcomings in the data collected by Lazarus et al. (2009) with regard to investigating mechanisms underlying the observed

macroevolutionary pattern, I look toward groups in which evolutionary relatedness has been well documented. Fortunately, a small minority of radiolarian genera have been revised or created anew based on careful stratigraphic observation linking one form to another. These genera, or in some cases small groups of genera, constitute phyletic or anagenetic lineages, along which species are segments (more or less arbitrarily) defined by morphological thresholds (Kling 1978; Lazarus 2005).

Comprehensively answering the question of how the Cenozoic change in radiolarian silicification occurred is a large problem requiring substantial taxonomic and phylogenetic work; simply establishing the phylogenetic relationships among enough of the taxa to make an authoritative statement about the assemblage as a whole will require long-term collaborative effort across the discipline. However, we can determine whether anagenetic change in silicification has occurred at all and, thus, whether anagenetic changes *could* account for the macroevolutionary pattern observed.

Operating in isolation, each of the two major processes—selection within species or selection between species—predicts different patterns of silicification within lineages. If the assemblage-level silicification trend is the result of macroevolutionary processes operating above the species level alone, that is, due to a mode of evolution consistent with the strict model of punctuated equilibria, we would expect to see no anagenetic silicification changes within radiolarian lineages. If, on the other hand, the assemblage-level trend resulted from selection for less-silicified morphologies operating on species directly, we would expect to see directional change commonly, if not universally, within anagenetic lineages.

5.2.3 Well-Described Radiolarian Lineages

Only a relatively small proportion of the total known radiolarian species diversity (around 15,000 according to Suzuki and Aita 2011) has been biostratigraphically defined within this framework of anagenetic or phyletic lineages; of these, only a small number proved to be of utility to the present study.

Because the assemblage-level signal is most pronounced in low latitudes (Lazarus et al. 2009), lineages confined to higher latitudes were not considered. Furthermore, some radiolarian test morphologies simply do not lend themselves to measuring the degree of silicification—for example, in taxa consisting mostly of spines, like *Dorcadospyris* (Sanfilippo et al. 1985), it is difficult to determine a cell volume of which a proportion of test silica could be quantified. Similarly, because the degree of silicification in spongiose fabric (three-dimensional sponge-like test porosity) is not readily measurable by transmitted light microscopy, largely spongiose taxa like *Lithocyclia* (Sanfilippo et al. 1985) were also avoided. Lineages spanning short intervals or intervals where no change in silicification is seen at the assemblage level (e.g. *Phormocyrtis,* Foreman 1973) were similarly not chosen. A further disqualifying factor for some lineages was low abundance; for example, members of the *Artophormis* lineage (Sanfilippo et al. 1985) largely occur as single specimens if they are present at all, and then often as fragments, precluding the collection of significant sample sizes through a time series.

In a few cases, for example the cluster of species related to *Lophocyrtis* (Sanfilippo 1990) or the Pterocorythidae (Sanfilippo and Riedel 1992), the evolutionary relationships among whole sets of anagenetic lineages have been traced out biostratigraphically, documenting the cladogenetic events linking them. In the case of the *Lophocyrtis* clade, although its long range would be ideal for the study at hand, the large number of species involved proved to be beyond the scope of an exploratory study.

After applying the filters of the criteria described above, three lineages where selected for which the anagenetic sequence of forms has been well-described and broadly accepted in the literature: *Stichocorys*, *Didymocyrtis*, and *Centrobotrys*.

Stichocorys

Classified within the nasselarian family Theoperidae (featuring a small, spherical, poreless or almost poreless cephalis with a reduced internal spicule, Riedel

1967), the genus *Stichocorys* is defined as follows:

> "Cenozoic (perhaps only Neogene), multi-segmented theoperids, in which the first three or four segments constitute a conical upper portion of the shell and the subsequent segments (narrower than the greatest width of the conical portion) constitute a cylindrical lower portion. Apical horn small, simple." (Sanfilippo and Riedel 1970)

The *Stichocorys* lineage (see Fig. 5.2) as considered in this study consists of three described species: *S. delmontensis, S. wolffii, and S. peregrina*. The anagenetic relationship of the two major species, *S. delmontensis* (the ancestor) and *S. peregrina* (the descendant), as well as their phylogenetic context within the closed Theoperids of the Cenozoic, are described in Sanfilippo and Riedel (1970). *S. delmontensis* includes those forms in which the conical part of the test consists of the first three segments; it ranges from the middle Early to Late Miocene (Sanfilippo et al. 1985). *S. peregrina* includes those forms in which the conical part of the test consists of the first four segments; its published range is from the Early to the Middle Pliocene (Sanfilippo et al. 1985). *S. wolffii* differs from *S. delmontensis* mainly in having a practically poreless thorax (six or fewer pores in the visible half of the segment). Sanfilippo et al. (1985) consider *S. wolffii* not to be a "good 'biological' species" but rather a morphological variant of *S. delmontensis*, although morphological differentiation beyond the thoracic porelessness is described at the end of its range by Riedel and Sanfilippo (1978).

References to taxonomy and figures for each of the species are given in Sanfilippo et al. (1985).

Didymocyrtis

The genus *Didymocyrtis* is classified within the spumellarian family Coccodiscidae (having discoidal or ellipsoidal shells enclosing single or double medullary shells, Sanfilippo and Riedel 1980) and is defined as follows:

> "Ellipsoidal cortical shell equatorially constricted in all but the earliest forms … Extra-cortical caps, when present, never more than

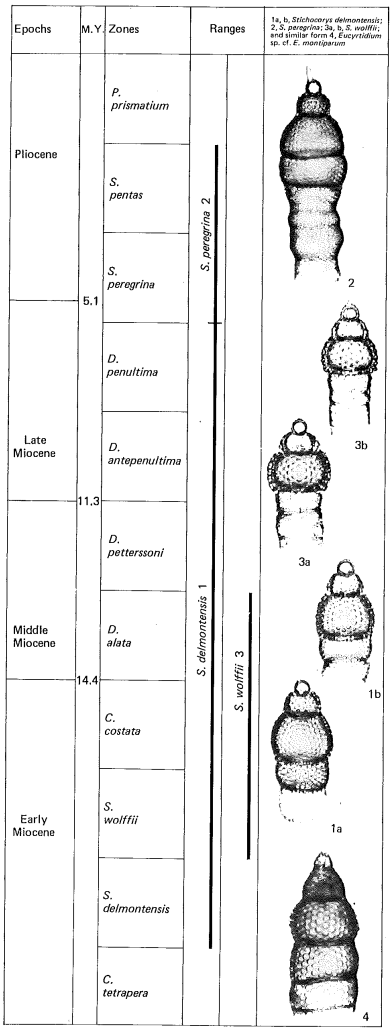


Figure 5.2: Figure reproduced from Sanfilippo et al. (1985) showing morphology and range of the three species in the *Stichocorys* lineage.

> two or three on each pole. Outer medullary shell commonly lenticular. Includes *D. prismatica* and *D. tetrathalamus*, [and] all the members of the evolutionary lineage between those two…"
> (Sanfilippo and Riedel 1980)

The phylogenetic relationships of the lineage, shown in Figure 5.3, were described by Kling (1978), though the generic names were subsequently revised by Sanfilippo and Riedel (1980). The lineage begins with the genus *Lithocyclia* in the Eocene, but those species are not considered here due to the difficulties of measuring silicification in spongiose texture described above. The lineage continues through the *Didymocyrtis* species *D. prismatica, D. violina, D. mammifera, D. laticonus, D. antepenultima, D. penultima, D. avita, and D. tetrathalamus* from the latest Oligocene through the Quaternary. This lineage includes a cladogenetic event giving rise to the *Diartus* lineage; however, again due to the abundance of spongiose texture in its test morphology, that branch was not included in this study. Further references and detailed descriptions for each of the species can be found in Sanfilippo et al. (1985).

Centrobotrys

The genus *Centrobotrys* of the nasselarian family Cannobotryids was first defined by (Petrushevskaya 1965) for the species *C. thermophila*. That species was subsequently found to be the extant member of a lineage (see Fig. 5.4) stretching back to the Early Oligocene species *C. gravida* (Moore 1971) via the mid-Oligocene intermediate species *C. petrushevskayae* (Sanfilippo and Riedel 1973). Though widely accepted in the literature, this lineage has received less attention than the two described above (D. Lazarus, pers. comm.). Detailed descriptions of the three species are provided in Sanfilippo et al. (1985).

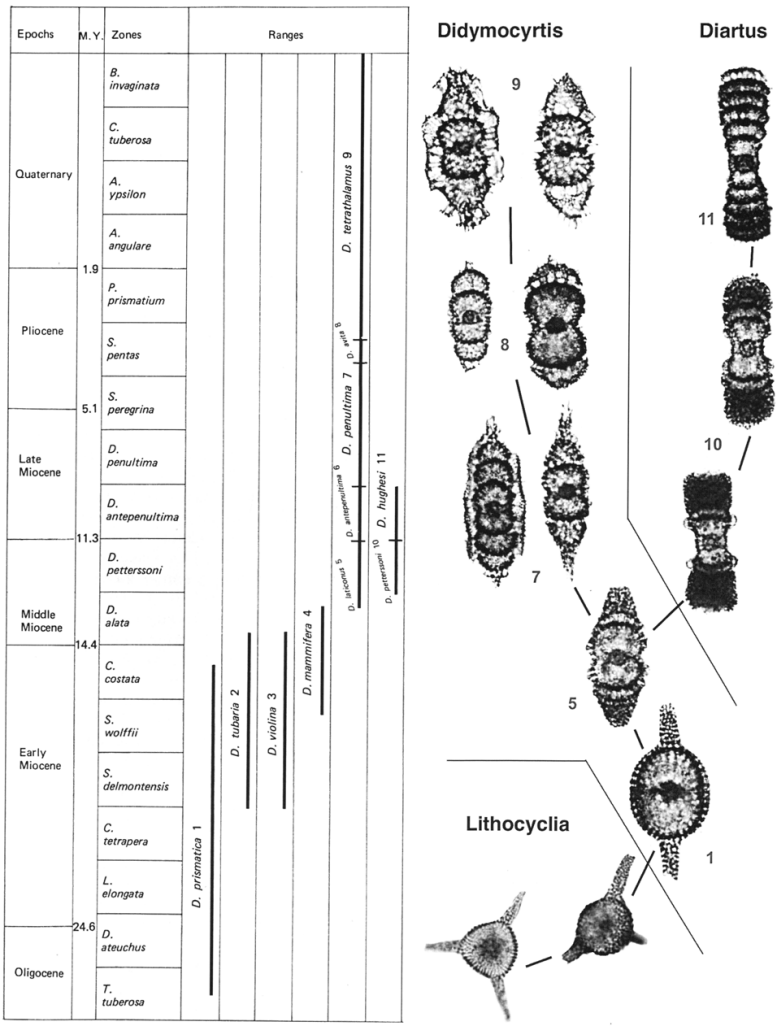


Figure 5.3: Figure reproduced from Lazarus (2005) showing morphology and range of the species in the *Didymocyrtis* lineage. Neither the *Diartus* branch nor the *Lithocyclia* portion of this clade were considered in this study.

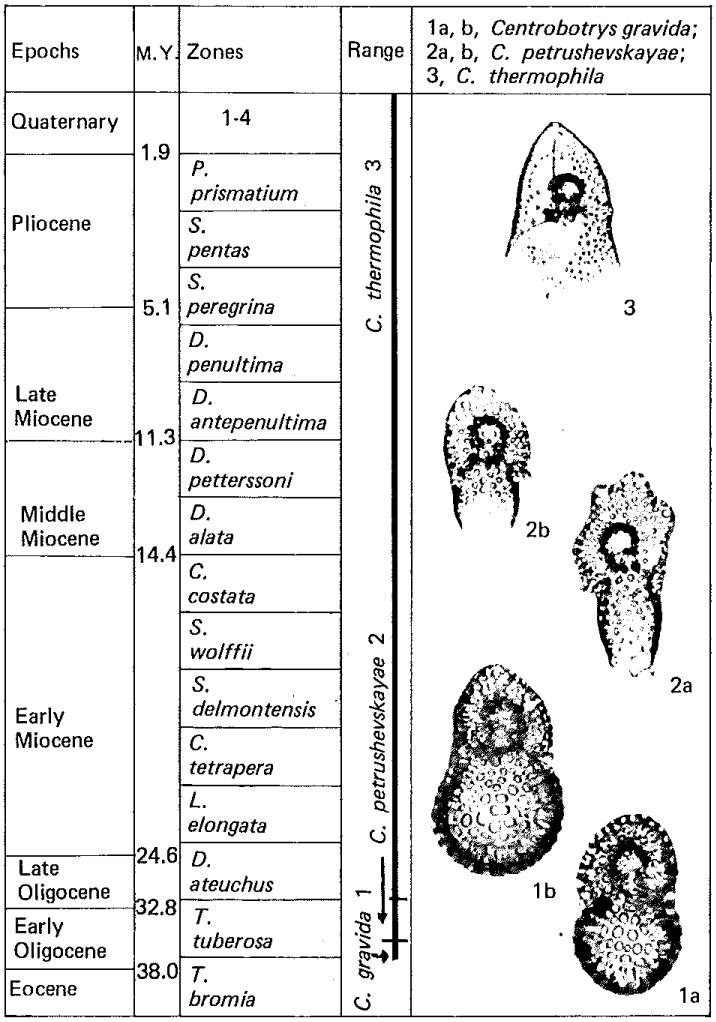


Figure 5.4: Figure reproduced from Sanfilippo et al. (1985) showing morphology and range of the three species in the *Centrobotrys* lineage.

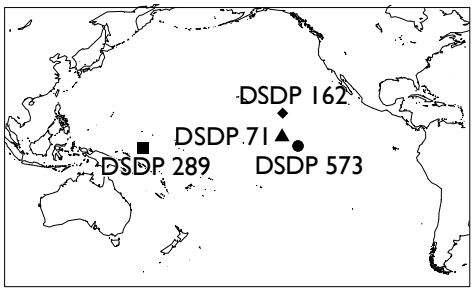


Figure 5.5: Map showing the location of DSDP drill sites from which samples used in this study were drawn.

5.3 Materials & Methods

5.3.1 Samples Used

The Integrated Ocean Drilling Program (IODP), the descendant of the Deep Sea and Ocean Drilling Programs (DSDP and ODP), maintains Micropaleontological Reference Centers (MRCs) around the world. These centers house collections of prepared slides of microfossils from many of the drill cores recovered since the initiation of ocean drilling in 1968, including many thousands of radiolarian slides. A database of MRC samples [1] contains curatorial information, including the depth below seafloor from which the samples were taken, but no data on sample age or taxonomic composition. In order to find samples likely to contain species of interest, I searched the *Neptune* database (Lazarus 1994; Spencer-Cervato 1999)—a separate compilation of tens of thousands of records of occurrences of microfossil species from DSDP and ODP publications—to identify boreholes and depth intervals from which the species of interest have been reported. I then used the MRC database to locate prepared slides from those boreholes and depth intervals.

During visits to two MRCs, at the Smithsonian Institution in Washington, DC, and the Museum für Naturkunde in Berlin, I inspected hundreds of radiolarian slides selected in this way to determine which contained the species of interest in

[1] Accessed at: iodp.tamu.edu/curation/mrc/MRC_database.txt

sufficient numbers. This was necessary since it is far from given that representatives of a radiolarian species will be present in every sample between its first and last appearance, even if the species is quite abundant when present; the composition of an assemblage can vary quite considerably from one sample to the next in a stratigraphic succession. A single drill site (DSDP-573) yielded time series for both the *Stichocorys* and *Didymocyrtis* lineages. For the *Centrobotrys* lineage, no single site provided a continuous record and so samples from multiple sites were used; to limit biogeographic bias, samples were chosen from the same region (the tropical Pacific).

A map showing the locations of drill sites used in this study is provided in Figure 5.5.

Sample Age Determination

The geological ages of the samples drawn were estimated by bracketing each sample between the nearest samples above and below in the *Neptune* database. *Neptune* ages are reported in terms of the Cenozoic timescale of Berggren et al. (1995); the same timescale is used here. I did not construct new age models for the relevant boreholes because the spacing between the bracketing ages from *Neptune* is generally much less than the spacing between the samples drawn for this study, and because the precise age of each sample is not expected to bear significantly on the outcomes observed. Nonetheless, it is worth noting that the age models used in deriving sample ages are in some cases not made explicit in the *Neptune* database. In these cases it appears that there is no record of which age model was used, nor how it was constructed (D. Lazarus, pers. comm.).

5.3.2 Method of Data Collection

A total of 6790 measurements of 44 morphological parameters (lengths, widths, thicknesses, and porosities on various test parts) were made using a transmitted light microscope with an attached digital camera connected to a computer running a custom software setup *(RadData)* for image analysis and storing

measurements. This measurement setup and its software components are summarized diagrammatically in Fig. 5.6. The measurements obtained in this way were then used in conjunction with geometric models of the relevant species (consisting of combinations of cones, spheres, cylinders) to calculate a silicification percentage for each measured specimen, i.e. the volume of silica as a proportion of the volume enclosing the test.

Microscopy

Prepared radiolarian slides were imaged on a Leitz Orthoplan microscope under brightfield illumination using the following objectives: Olympus DPLAN 4x 0.1 and 10x 0.25, Leitz PL FLUOTAR 25x 0.6 and 40x 0.7, and Nikon PH3DL 60x 0.7 LWD. High magnification and high numerical aperture objectives were required in order to resolve the µm-scale test thicknesses. However, the preparation of the MRC slides used precluded the use of conventional objectives. Because these slides were prepared with taxonomic work in mind, for which high magnification is not usually required, many contain specimens beneath a thick, ~1 mm layer of Canada balsam under the standard 0.17 mm coverslip. Bringing such specimens into focus requires a free working distance of >1 mm, necessitating the use of a long working distance objective.

Digital images were captured using a Canon EOS Rebel T1i digital camera connected to the photo port of the microscope using a Diagnostic Instruments PA1-35A adapter. The camera was controlled via a USB connection through Canon's EOS Utility software running on a MacBook Pro computer. Custom software written using the *R* programming language (R Development Core Team 2011) displayed an illustrated list of the relevant parameters to be measured for the species at hand (see Fig. 5.7). Each digital image acquired was opened in the image analysis software ImageJ (Abràmoff et al. 2004). Measurements of test parameters (lengths, widths, thicknesses, and porosities on various test parts) were made in ImageJ and automatically recorded to a text file using custom ImageJ macros. These measurements were then automatically read into the *R*

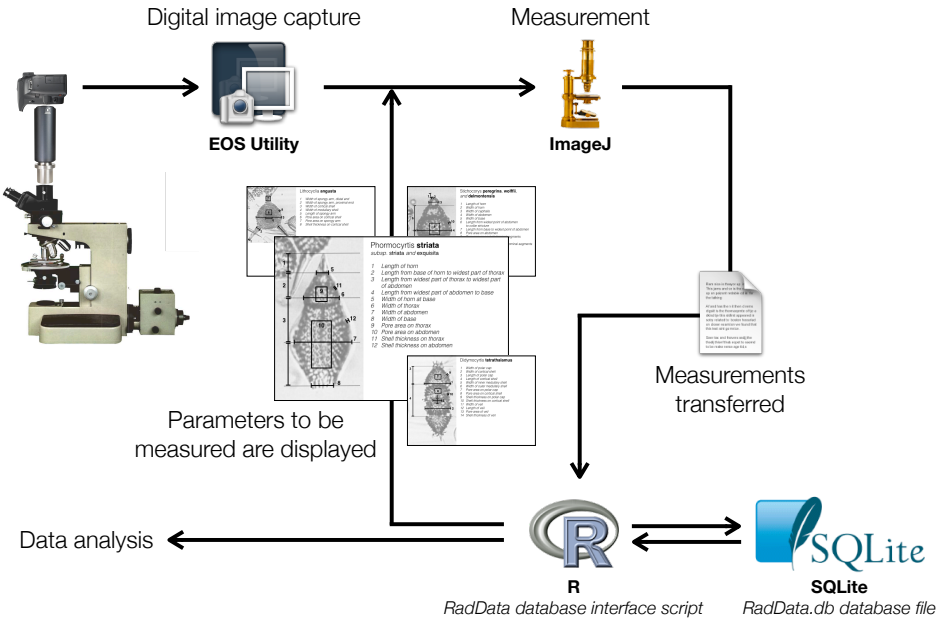


Figure 5.6: The setup used for making, storing, and analyzing measurements. Prepared slides are examined under brightfield illumination on the microscope and images are digitized with a Canon digital camera controlled by Canon's EOS Utility software, set to automatically open the acquired image in ImageJ, where measurements are made. A custom macro in ImageJ saves these measurements to a file, from where they are read by custom software in R and saved to a SQLite database. This database can, in turn, be interrogated from R for data analysis.

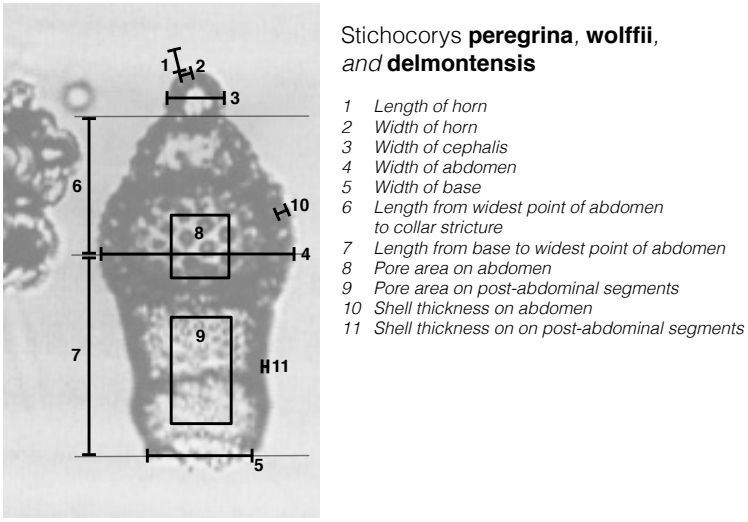


Figure 5.7: Example of a prompt screen from the interface to the RadData database. When the user has encountered a specimen to be measured and identified it taxonomically to the program, a screen like this one is displayed to prompt the measurement of the relevant morphological features in the correct order.

software and stored in a database file using an *R* implementation of the relational database management system SQLite.

The design of the *R* software and the measurement workflow surrounding it is summarized schematically in Figure 5.8. The source code for the custom software and macros written for the purposes of this study is provided in Appendix G.

Advantages of Relational Databases

Prior experience in collecting radiolarian datasets with many thousands of measurements (Lazarus et al. 2009) highlighted two major shortcomings of using spreadsheets (such as Microsoft Excel) to collect and analyze large datasets. First, analyzing data from such large spreadsheets is cumbersome, requiring laborious and repetitive tasks such as the manual selection of specimens that match certain criteria, resulting in large and complex files with many sheets linked by complex formulae. This is particularly true when restructuring of the

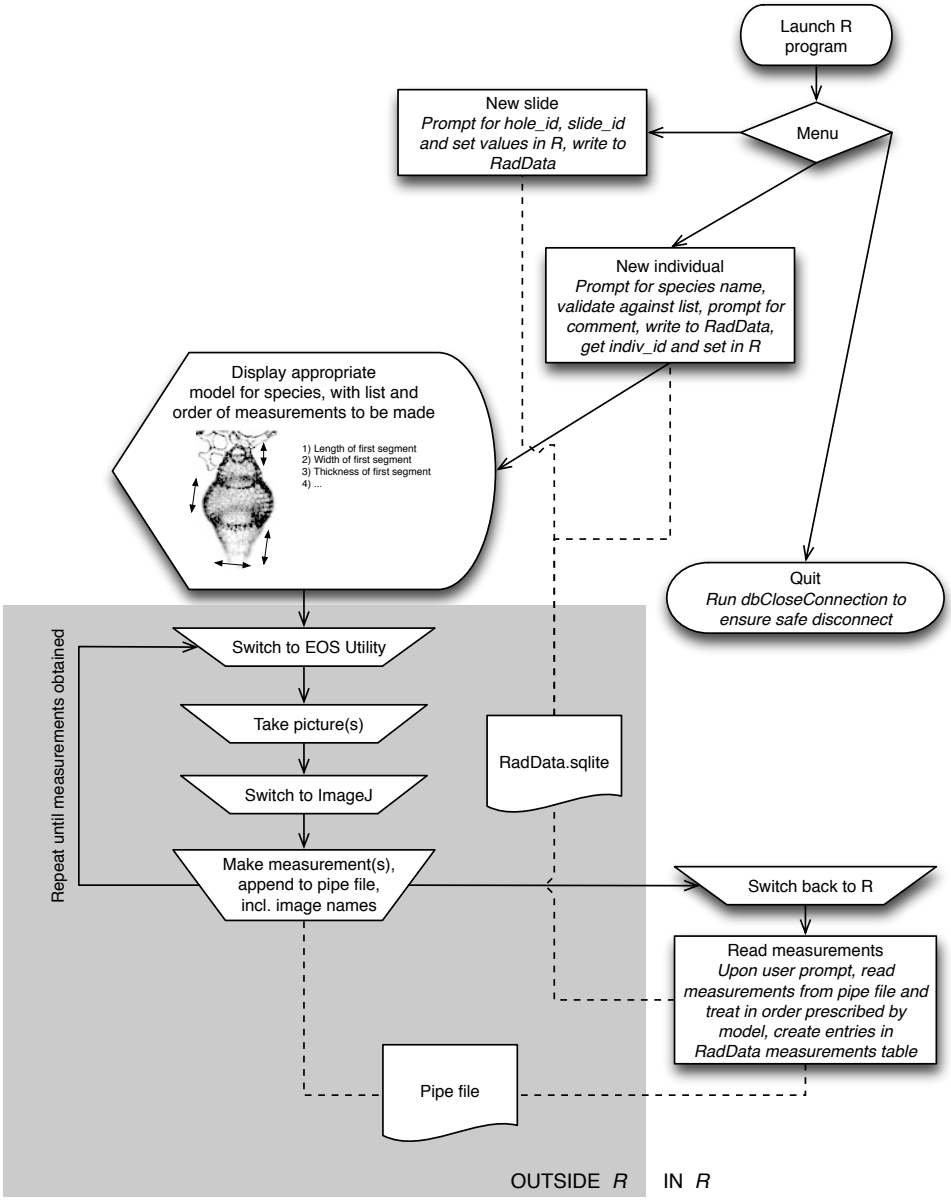


Figure 5.8: This schematic chart shows the design of the R software to make and store measurements, in the context of the measurement process.

data is required—such as separating data by geography or taxonomy. Second, the files can become unstable and corrupted.

The use of relational databases to store such data alleviates these problems, but also confers a number of other advantages. In a spreadsheet, data are held in a single table, with each row often consisting of pieces of information describing many different entities. For example, each row of the spreadsheet containing the radiolarian data collected for the Lazarus et al. (2009) study contains fields with data describing a borehole (the hole IDs, latitudes and longitudes), a slide (sample depth, sample age), and an individual radiolarian (taxonomic classification, measurements of test morphology). In a relational database, by contrast, separate tables collect information related to only one entity each; these tables, in turn, are linked by a strictly defined organization that reflects the underlying structure of the data (a design process termed "normalization"). The *RadData* database constructed for this study consists of separate tables containing the data describing boreholes, slides from those holes, individual specimens from those slides, and measurements on those individuals separately.

Relational databases thus result in more efficient, smaller, and more stable files. But, more importantly, they allow data to be brought into new relationships easily, especially when those relationships were not anticipated at the time the data were collected. This is achieved by querying the database using algebraic expressions based on operators similar to those familiar from set theory (e.g. union, intersect), instantaneously producing new arrangements and conjunctions of the chosen data from the separate tables.

RadData database

The schema for the *RadData* database, i.e. its organizational blueprint, is shown in Figure 5.9. Each of the four boxes represents a table, which can be thought of as a spreadsheet with rows and columns; the fields or column names are shown in each box. The arrows show how the tables are logically linked by shared information between tables; for example, the *Slides* table contains a field with the

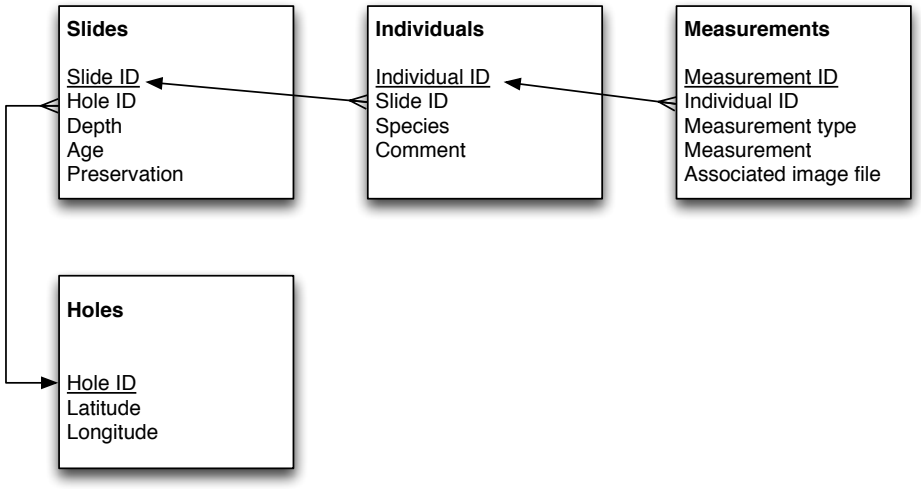


Figure 5.9: Database schema (organizational blueprint) of the *RadData* database used to store measurements of radiolarian test morphology. Following the principles of relational database design, each table of the database groups attributes of a specific entity. For example, the *Slides* table contains information about the borehole depth from which the slide was obtained and the preservation of the slide, but not about the longitude and latitude—since those are attributes more generally of the borehole.

Hole ID identifying the borehole from which the sample was taken. This links entries in the *Slides* table to the *Holes* table. This design is similar to that used in the *Neptune* database (Lazarus 1994).

The source code for creating the *RadData* database is provided in Appendix G.1.

Geometric Models of Species

In order to calculate percent silicification, the measurements of test morphology were used as inputs to models of test morphology constructed from combinations of simple, hollow geometric volumes like cones, spheres, cylinders, and portions thereof. Test porosity was taken into account by multiplying the volumes obtained by the complement of the relevant measured value of surface pore area (ranging from zero to one). An example model, used for the species of

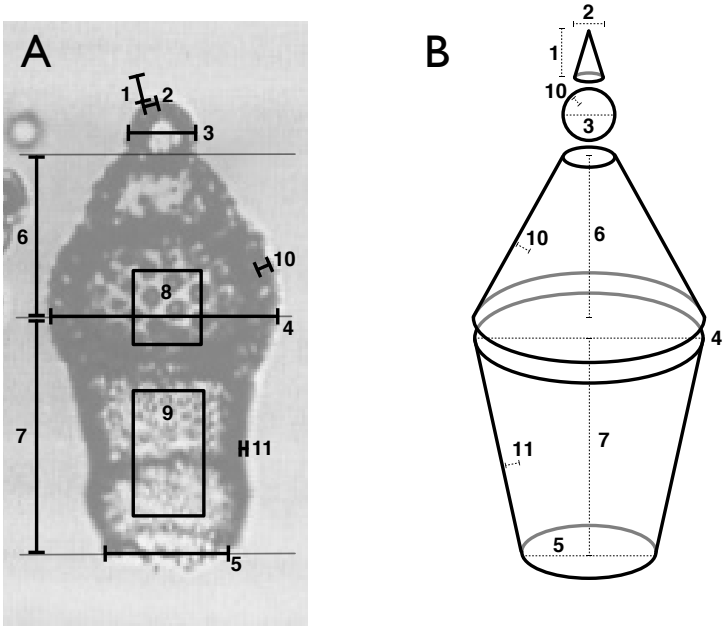


Figure 5.10: The geometric model used to calculate silicification from morphometric measurements for the species *Stichocorys delmontensis* and *Stichocorys peregrina*. A, measurements made (see Fig. 5.7 for explanation of labels). B, geometric shapes used in calculations. The cone representing the apical horn is solid, while the remaining forms are hollow, their silica volumes calculated by subtracting a similar volume smaller in size by the corresponding measured test thickness. Additionally, the cone frusta representing the postcephalic segments are multiplied by the fraction of surface area not occupied by pores to calculate their contribution to test silica volume. Silicification percentage is calculated as a proportion of the total volume (in this case, the sum of the outer volumes of the cone, sphere, and two cone frusta).

the *Stichocorys* lineages, is shown in Figure 5.10. A total of 8 morphological models were used in the results presented below; the source code for the software written to perform these calculations is provided in Appendix G.4.

5.3.3 Model Fitting

It has long been appreciated that trends in paleontological time series can in principle result from undirected, stochastic processes like random walks (e.g. Raup and Gould 1974) as well as from directional selection. From the array of

tests that have been developed over the years to evaluate such hypotheses for paleontological time series data, I applied the likelihood-based approach of Hunt (2006) that provides a statistical comparison of competing models, as opposed to one based on the rejection of null hypotheses.

The analysis was carried out using the *paleoTS* package (Hunt 2012) for the *R* programming environment. This method analyzes sample means, variances, and ages in order to compute Akaike weights for three explicit evolutionary models (measures of the relative plausibility of models given the data at hand). The three models are an unbiased random walk, where the mean step size from one time to the next is zero (representing non-directional evolution, i.e. Brownian motion), a generalized random walk, where the mean step size is not zero (representing directional evolution), and evolution around an optimal phenotype, where the step size and direction at each time step depends on the ancestral state relative to that optimum (representing stasis). The computation of the adjusted Akaike Information Criterion (AICc) at the heart of this approach takes the complexity of the model into account; models are effectively penalized for having a greater number of parameters (K).

5.4 Results

5.4.1 Sample Size

In order to determine the sample size required to adequately characterize the percent silicification of a particular lineage in a particular slide, a relatively large number of specimens (100) of a single lineage *(Stichocorys)* were measured on one slide. To obtain an estimate of the precision in average silicification values at different sample sizes, the calculated silicification percentages were randomly reordered 500 times and the average calculated at each sample size from 1 to 100 (see Fig. 5.11).

The resulting distribution of average values narrows rapidly up to a sample size of 20-30 individuals; above that sample size, returns in precision diminish for

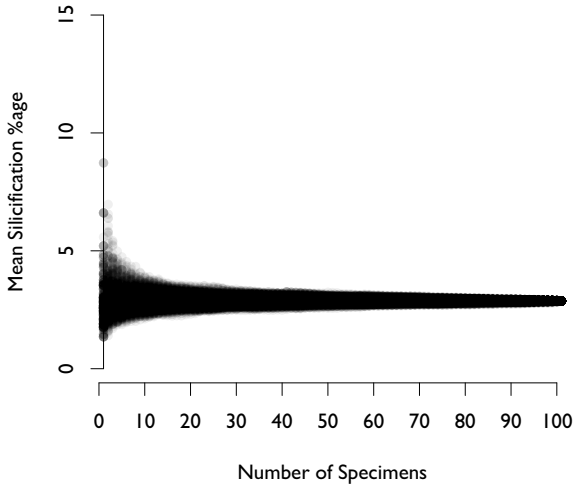


Figure 5.11: Mean silicification (in % of test volume) plotted against sample size for the *Stichocorys* lineage (*Stichocorys peregrina*) in slide DSDP-85-573-7-6,41-48. The 100 specimens (individual radiolarians) measured on this slide were randomly ordered and mean values calculated for increasing numbers of specimens included. This process was repeated 500 times to obtain an envelope roughly showing the 'precision' of an average value at a given sample size relative to the average at a sample size of 100. The returns in precision from adding additional specimens diminish beyond 20-30 individuals; thus a target sample size of 25 specimens per slide was chosen.

further sample size increases. A target sample size of 25 specimens was chosen, though it is noted that not every slide examined contained 25 or more specimens of each lineage of interest. *Centrobotrys* species are relatively rare and no slide examined reached this threshold.

5.4.2 Secular Changes in Silicification in Radiolarian Lineages

Stichocorys

Histograms showing the distributions for the degree of silicification of the two species in this lineage, *S. delmontensis* (the ancestor) and *S. peregrina* (the

descendant), as well as the morphological variant of the ancestor, *S. wolffii*, are shown in Figure 5.12. The distributions are clearly different, with the descendant species centered on a lower percent silicification (mean value 3.62%) than the ancestor (mean value 6.37%). Indeed, a two-sample T-test suggests a rejection of a null hypothesis of equal means with a p-value of $< 2.2 \times 10^{-16}$. These results suggest that the average degree of silicification decreased over the evolution from the ancestral to the descendant species. *S. wolffii*, however, clearly has a higher degree of silicification than either of the main species.

A different view of the same data is presented in Figure 5.13, which shows the results plotted through time, with average values calculated for each slide. The same general pattern of a net decline through time is seen, with a slight early to mid-Miocene rise from around 6–9%, where *S. wolffii* is present, before a steady decline through the second half of the Miocene to Pliocene values around 3%. The secular decline in mean silicification values in the second half of this time series is accompanied by a decrease in the maximum and minimum values.

Table 5.1: Results of Akaike Information Criterion (AIC) analysis using software by Hunt (2012), comparing the relative goodness of fit for three evolutionary models to the silicification for *Stichocorys* silicification (Fig. 5.13). Models are generalized random walk, i.e. directional change (GRW), unbiased random walk, i.e. non-directional change (URW), and stasis. Results shown are the log-likelihood (higher is preferred), number of model parameters, K, AICc (a small-sample-corrected version of AIC; lower is preferred), and Akaike weight (higher is preferred). Models with greater than minimal support (> 0.05) shown in bold.

	logL	K	AICc	Akaike weight
GRW	−21.24	2	47.82	**0.22**
URW	-21.46	1	45.31	**0.78**
Stasis	-25.91	2	57.15	0.00

The results of the model selection analysis shown in Table 5.1 suggest that the non-directional evolution model (unbiased random walk) is the best of the three models given the observed data. The results also afford some support for the directional evolution model, but no support for the stasis model.

I note that the trend observed could be explained by two separate episodes of directional evolution, a directional rise until ~15 Ma, followed by a directional

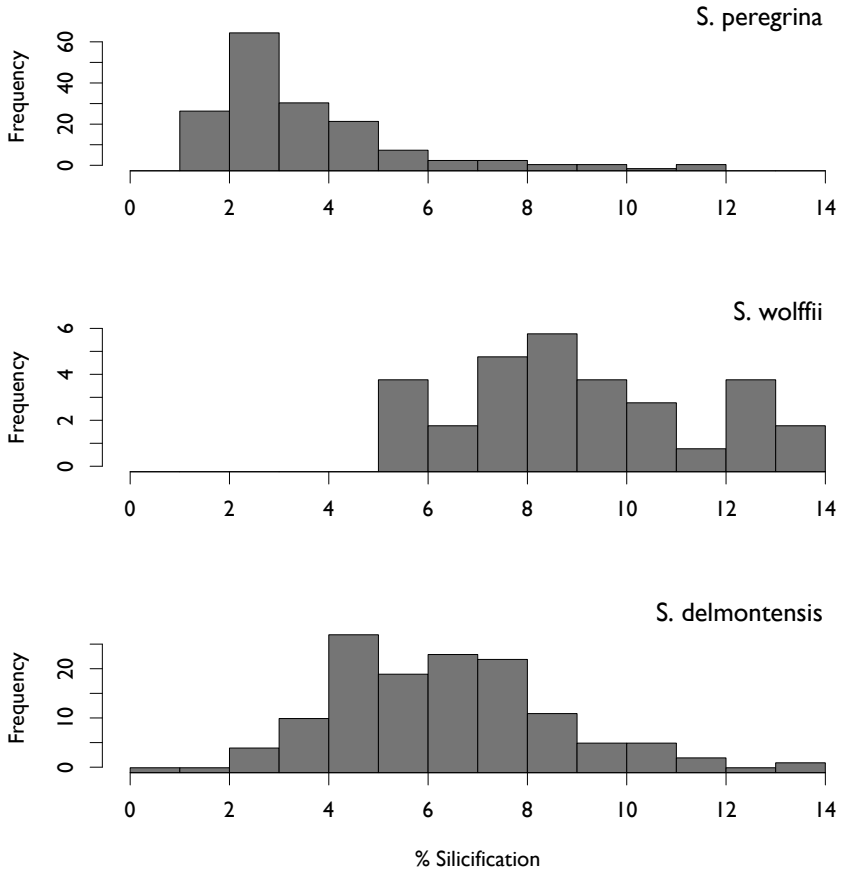


Figure 5.12: Histograms of the degree of silicification (in % of test volume) for the species in the *Stichocorys* lineage in the equatorial Pacific. The descendant species, *S. peregrina*, shows a distribution of silicification values that is shifted toward lower values than the ancestor; the morphological variant, *S. wolffii*, is more heavily silicified.

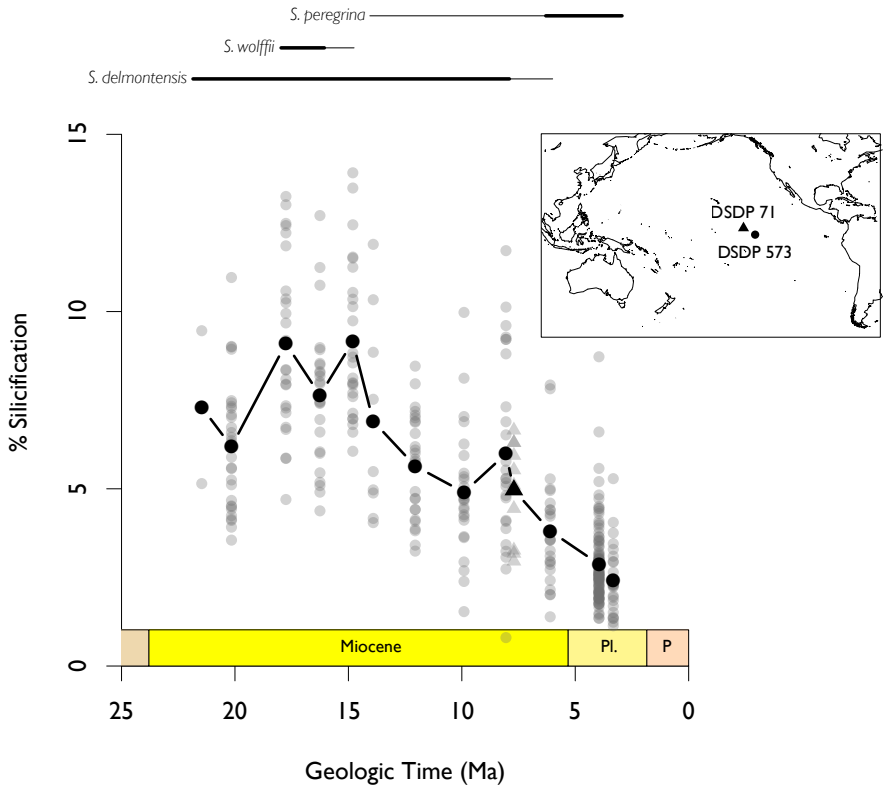


Figure 5.13: Degree of silicification (in % of test volume) through time for the *Stichocorys* lineage in the equatorial Pacific. Translucent grey points represent values for individual specimens; solid black points represent sample averages. Plot symbol indicates DSDP drill site (see inset map). Bars above plot indicate temporal range of species in the lineage, as encountered in this study.

fall. The generalized statistical technique to extend model fitting to allow for shifting evolutionary dynamics (i.e. including multiple segments with different models for each) has been developed (Hunt 2008). I do not apply this approach here, however, since segments shorter than about five samples can produce spurious results (Hunt 2008, p. 364).

Didymocyrtis

Figure 5.14 shows histograms of the degree of silicification for seven species in the *Didymocyrtis* lineage, arranged in stratigraphic order. Examining first the beginning and end-points of this anagenetic lineage, we see a first-order similarity to the results for *Stichocorys*: the youngest species in the lineage, *Didymocyrtis tetrathalamus* is less silicified (mean value 2.17%) than the oldest species, *Didymocyrtis prismatica* (mean value 8.84%). This difference in means is also highly significant in a T-test comparing the values in those two species (a p-value of 3.3×10^{-10}).

Much as for *Stichocorys*, examining the silicification distributions of the intervening species along the lineage reveals a more complex pattern. There is a clear shift in the distributions to progressively more heavily silicified values until *Didymocyrtis laticonus*, where the trend reverses.

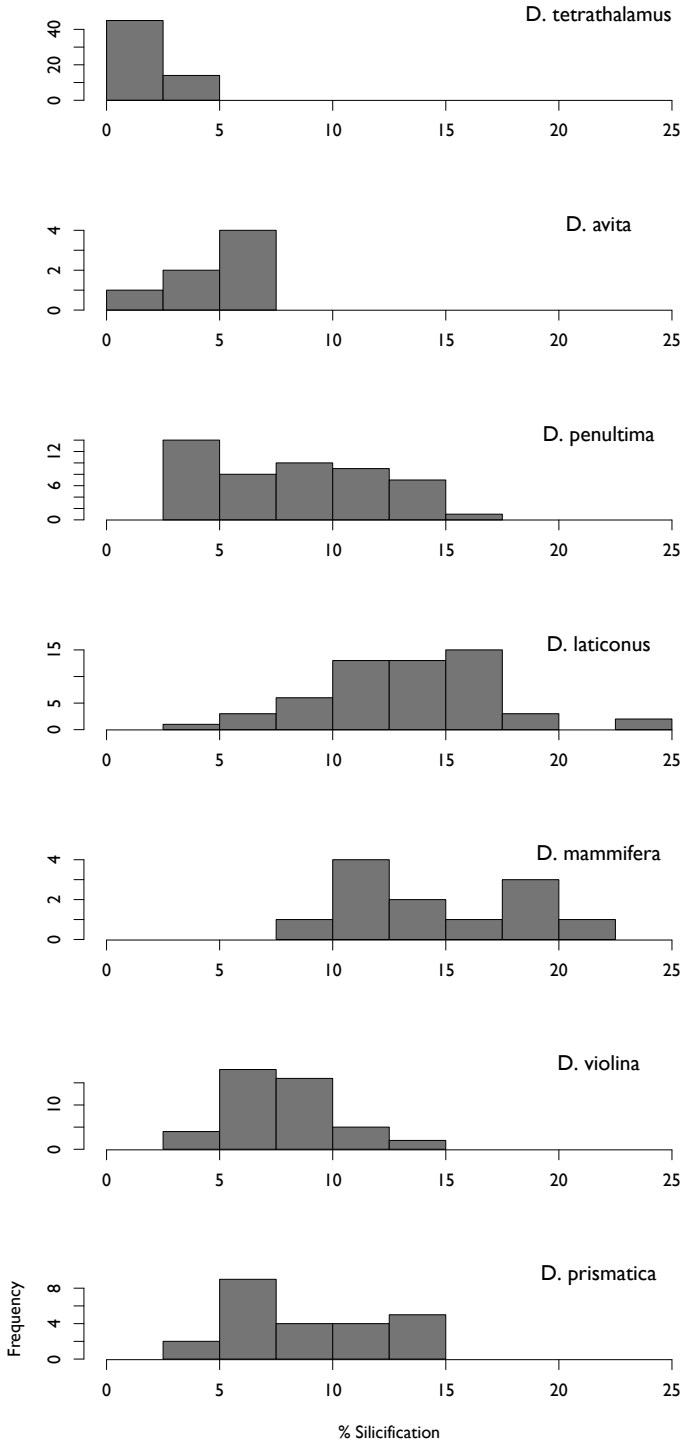
Figure 5.15 shows the same data on the degree of silicification in the *Didymocyrtis* lineage plotted through geologic time. Mean silicification values begin in the early Miocene at around 10%, drop slightly to around 7%, and then climb to a peak close to 15% at around 10 Ma, after which they fall sharply to end in the Pliocene at around 2%.

Although the trajectories of both *Stichocorys* and *Didymocyrtis* feature a rise and subsequent decline in the younger part of the time series, the onset of that decline is not synchronous; while the fall in silicification begins at around 15 Ma in *Stichocorys*, it does not begin until around 10 Ma in *Didymocyrtis*.

Although the beginning and end-points of this time series have averages that are significantly different, it is difficult to see a strong overall trend by visual

Figure 5.14 *(following page)*: Histograms of the degree of silicification (in % of test volume) for seven species in the *Didymocyrtis* lineage in the equatorial Pacific; arranged in stratigraphic order (youngest species on top). The distribution of silicification values appears to shift toward higher values in the middle of the lineage.

Figure 5.14: (continued)



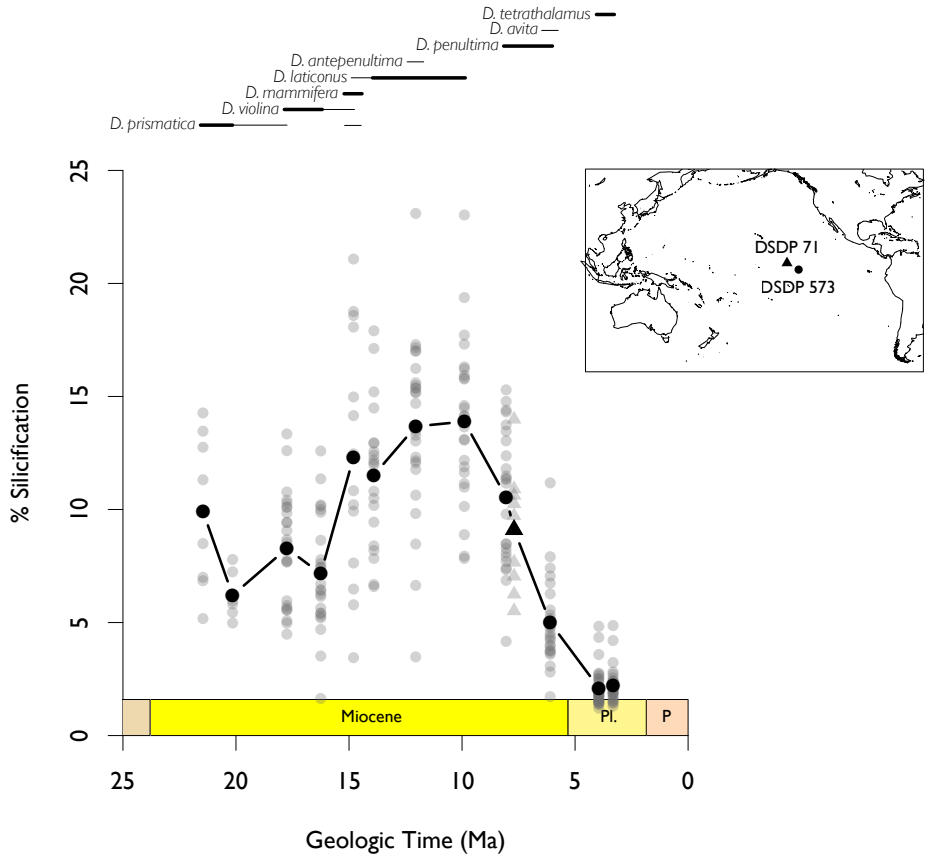


Figure 5.15: Degree of silicification (in % of test volume) through time for the *Didymocyrtis* lineage in the equatorial Pacific. Translucent grey points represent silicification percentage values for individual specimens; solid black points represent sample averages. Plot symbol indicates DSDP drill site (see inset map). Bars above plot indicate temporal range of species in the lineage, as encountered in this study.

inspection of this plot—in particular if the two Pliocene data points are disregarded. This makes a model analysis as carried out for *Stichocorys* above, pitting against one another models for directed evolution, random walk, and stasis, all the more pertinent.

In spite of showing a qualitatively different secular trajectory to that of the *Stichocorys* lineage, the model selection results for the *Didymocyrtis* lineage (Table 5.2) are very similar. There is no support for the stasis model, some limited support for the directional evolution model, but the bulk of the evidence appears to support the unbiased random walk model. Much as for *Stichocorys*, it is possible that the *Didymocyrtis* lineage is best described by two model segments with differing evolutionary dynamics. Crucially, however, the apparent shift point occurs at a different time in the two lineages (around 15 Ma in *Stichocorys* and 10 Ma *Didymocyrtis*).

Table 5.2: Results of Akaike Information Criterion (AIC) analysis using software by Hunt (2012), comparing the relative goodness of fit for three evolutionary models to the silicification for *Didymocyrtis* silicification (Fig. 5.15). Models are generalized random walk, i.e. directional change (GRW), unbiased random walk, i.e. non-directional change (URW), and stasis. Results shown are the log-likelihood (higher is preferred), number of model parameters, K, AICc (a small-sample-corrected version of AIC; lower is preferred), and Akaike weight (higher is preferred). Models with greater than minimal support (> 0.05) shown in bold.

	logL	K	AICc	Akaike weight
GRW	-28.20	2	61.74	**0.23**
URW	-28.46	1	59.32	**0.77**
Stasis	-33.43	2	72.20	0.00

Centrobotrys

Figure 5.16 shows histograms of the degree of silicification for the three species in the *Centrobotrys* lineage, again in stratigraphic order. The distributions do not appear distinct from one another; indeed, a T-test on the mean silicification percentages of *C. gravida* (8.43%) and *C. thermophila* (11.08%) returns a p-value of 0.13.

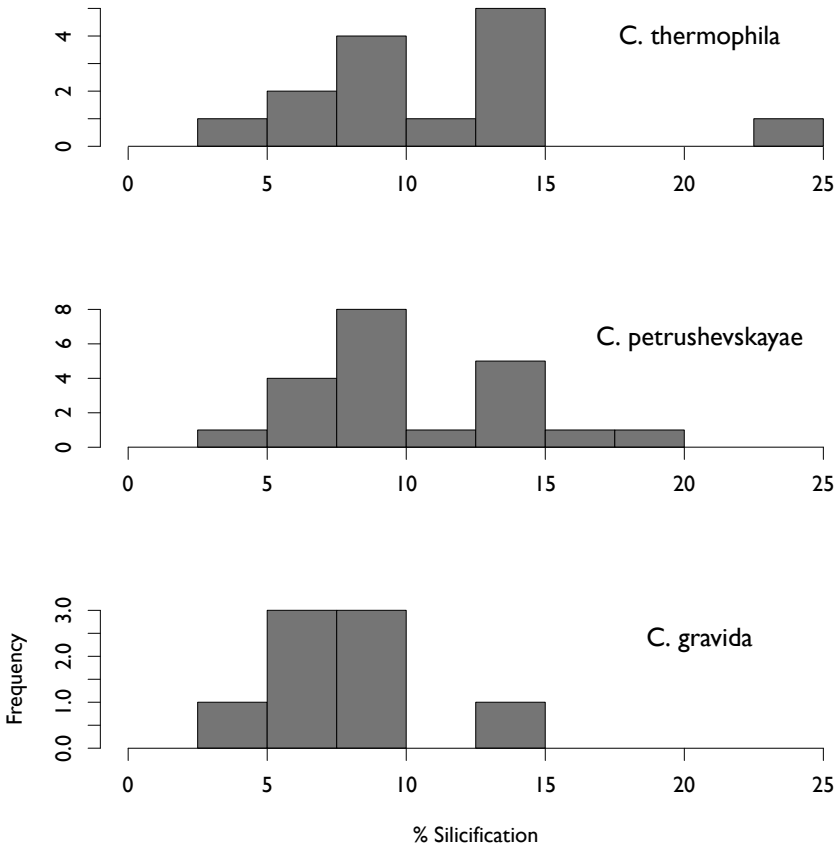


Figure 5.16: Histograms of the degree of silicification (in % of test volume) for seven species in the *Centrobotrys* lineage in the equatorial Pacific; arranged in stratigraphic order (youngest species on top). The distribution of silicification values do not appear to be different among species in this lineage.

Plotted through time (Fig. 5.17), the data tell much the same story: there is no discernible secular trend in silicification, nor much variation in the time series.

The stationary picture suggested by these results is confirmed by the model selection analysis (Table 5.3), which finds most support for the model describing stasis.

Table 5.3: Results of Akaike Information Criterion (AIC) analysis using software by Hunt (2012), comparing the relative goodness of fit for three evolutionary models to the silicification for *Centrobotrys* silicification (Fig. 5.17). Models are generalized random walk, i.e. directional change (GRW), unbiased random walk, i.e. non-directional change (URW), and stasis. Results shown are the log-likelihood (higher is preferred), number of model parameters, K, AICc (a small-sample-corrected version of AIC; lower is preferred), and Akaike weight (higher is preferred). Models with greater than minimal support (> 0.05) shown in bold.

	logL	K	AICc	Akaike weight
GRW	-18.03	2	44.05	0.02
URW	-18.03	1	39.06	**0.23**
Stasis	-14.37	2	36.74	**0.75**

Summary of Lineage Results

In summary, the three lineages studied show disparate results. In two lineages (*Stichocorys* and *Didymocyrtis*), there is a discernible net shift in silicification toward lighter values across the time series, while the third (*Centrobotrys*) shows no net change. The secular pattern is distinct in each, showing a decline following a slight rise in *Stichocorys*, a rise followed by a similar decline in *Didymocyrtis* (but with a different timing of its onset), and a stationary pattern in *Centrobotrys*. When the relative viability of three mathematical models of evolutionary modes is assessed using Akaike weights, the random walk model has the highest level of support in two of the lineages (*Stichocorys* and *Didymocyrtis*), while the model describing stasis is best supported in *Centrobotrys*. Directional evolution is thus not strongly supported in any of the three lineages.

Figure 5.18 shows the average silicification for each lineage through time compared to the average silicification for entire assemblages (Lazarus et al.

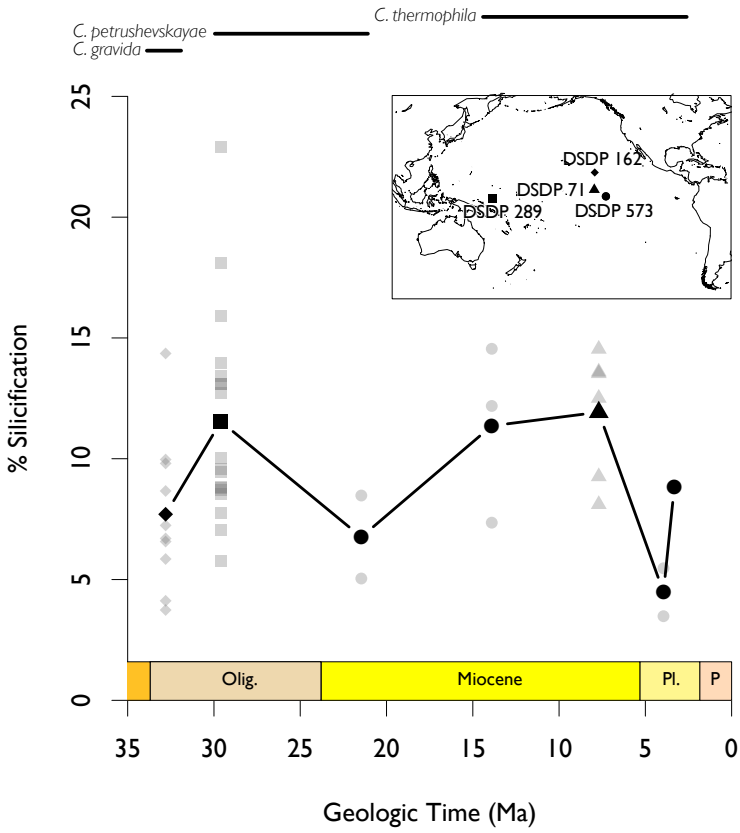


Figure 5.17: Degree of silicification (in % of test volume) through time for the *Centrobotrys* lineage in the equatorial Pacific. Translucent grey points represent silicification percentage values for individual specimens; solid black points represent sample averages. Plot symbol indicates DSDP drill site (see inset map). Bars above plot indicate temporal range of species in the lineage, as encountered in this study.

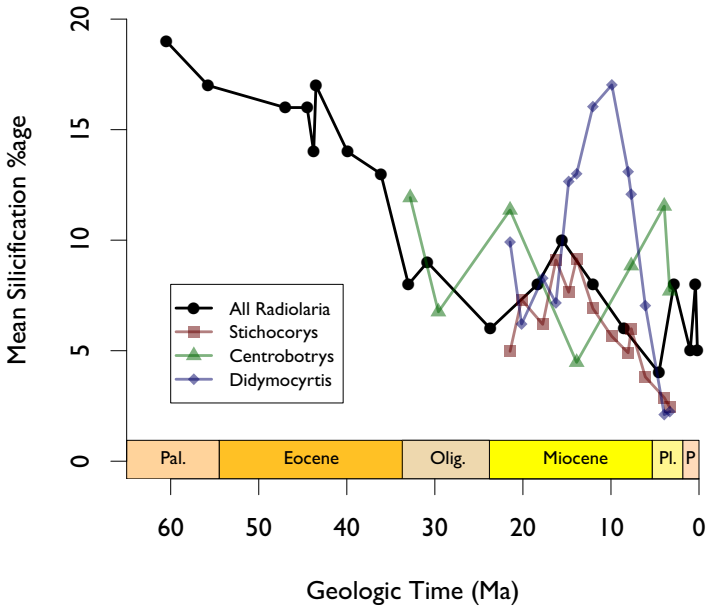


Figure 5.18: Percent silicification in the three lineages studied here (red, green, and blue lines) plotted in comparison to percent silicification of the whole radiolarian assemblage (black line, data from Lazarus et al. 2009).

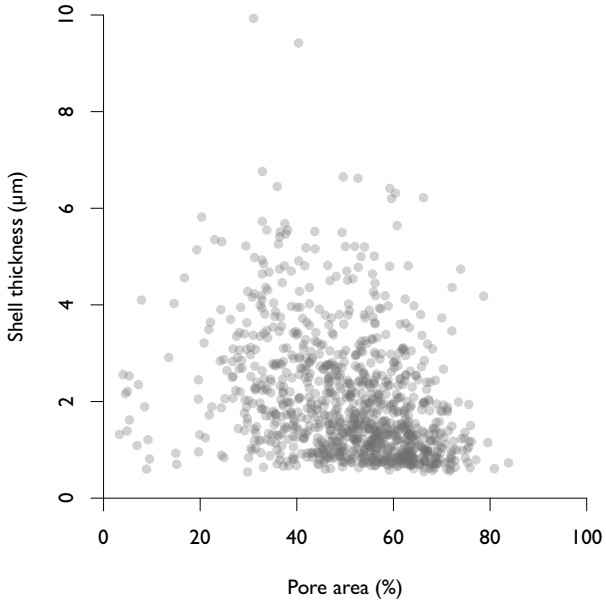


Figure 5.19: Test thickness plotted against porosity for all pairs of measurements made (including pairs of measurements on different parts of the same specimen, e.g. on the thorax and abdomen). There is no apparent relationship between test thickness and porosity, perhaps suggesting that the mechanical strength of the radiolarian frustule is less important than that of the diatoms.

2009). The results for *Stichocorys* track the whole-assemblage results closely, both in trajectory and absolute value. The *Centrobotrys* trend, on the other hand, shows more variability and no discernible relationship to the whole-assemblage results. The results for *Didymocyrtis* are quite different from the whole-assemblage results. While silicification values are comparable to the whole assemblage at the outset of the time series, they show a rise to a late-Miocene peak while the whole-assemblage values decline.

5.4.3 Relationship Between Test Thickness and Porosity

In a semiquantitative SEM survey of Mesozoic and Cenzoic diatoms Finkel and Kotrc (2010) found a relationship between frustule porosity and thickness in which thinner frustules were suggested to be less porous. In order to investigate

whether a similar relationship may exist in Cenozoic Radiolaria, test thickness was plotted against pore area from the data set collected here. No correlation was found between the two (see Fig. 5.19).

5.5 Discussion

The three lineages investigated do not vary in unison, but show distinct trajectories through time (Fig. 5.18). This observation is at odds with the coordinated trajectories we might expect to see if there was an overriding selective force acting at the species level upon all radiolarians. To a first order, the results do not support pervasive species-level anagenesis as a cause for the macroevolutionary pattern observed.

The lineages investigated are taxonomically distant, which may explain their disparate responses to the inferred decline in silica availability. *Stichocorys* are nasselarians while *Didymocyrtis* are spumellarians, separate clades at a high level in radiolarian molecular phylogenies (Suzuki and Aita 2011). Ecological research on living members of these clades has shown that they exhibit fundamentally different feeding strategies, closely related to their test morphology (Matsuoka 2007): multi-segmented nasselarians (like *Stichocorys*) capture relatively large prey, including ciliates and flagellates, using a prominent central pseudopodium and a cone of smaller pseuopodia that project from the opening of the test; solitary spumellarians (like *Didymocyrtis*) capture smaller prey, like microalgae and bacteria, using a large number of pseudopodia radiating from the test. This fact could explain differing trajectories in the framework of biomineralization costs and benefits (Knoll 2003), in which we can consider the evolution of the test as a balance between the costs of silica precipitation and the benefit derived from structuring the cytoplasm and providing mechanical support to pseudopodia (Anderson 1983). Two taxa with different factors on the benefits side of this equation might respond differently even when shifts on the cost side (e.g. by declining silicic acid concentrations) are the same for both.

In diatoms, the cost-benefit calculation for silica biomineralization has often

been explained by ascribing a mechanically protective function to the frustule, acting as armor to protect the cell from grazers (Hamm et al. 2003; Smetacek et al. 2004). In this context, the observation that diatom frustules may have become less porous as they became thiner-walled through Cenozoic time (Finkel and Kotrc 2010) can be understood as an evolutionary compensation to make a stronger frustule with less material. Based on the data collected in this study (Fig. 5.19) and the increase in porosity accompanying the decrease in test wall thickness reported from the whole-assemblage data (Lazarus et al. 2009), radiolarians do not show the same relationship. The explanation may lie in the rather different functional role played by the radiolarian test, alluded to above: while the diatom frustule might be analogically compared to a clamshell for its defensive function, the radiolarian test has been compared to the vertebrate skeleton and its supportive role (Anderson 1983). As such, differing functional constraints on the radiolarian test and the diatom frustule may have resulted in different evolutionary responses to declining silica levels. If the maintenance of compressive mechanical strength was not required in radiolarians, the cost-benefit balance may simply have shifted toward tests that were both thinner-walled and more porous.

None of the three radiolarian lineages studied here show silicification trajectories that are best fit by the evolutionary model for directional trends. This is consistent with the predictions of punctuated equilibria (Eldredge and Gould 1972) and the findings of the broad survey by Hunt (2007). Insofar as the absence of directional trends within lineages is representative of Cenozoic radiolarians as a whole, it suggests that the macroevolutionary pattern is driven by processes apparent above the species level, like species sorting and species selection, and not simply by anagenesis within lineages.

If Cenozoic radiolarian lineages mostly exhibit random-walk-like evolution in their degree of silicification, it is worth considering whether the assemblage-wide decrease in silicification resulted from macroevolutionary "diffusion," that is, an increase in variance from a bounded starting point (Gould 1988; McShea 1994). In that case we would expect to see maximum silicification values remain

constant while the minimum declined, causing a decrease in average silicification through an increase in variance. Such a pattern, however, is observed in neither the lineages examined here nor in the assemblage as a whole (Lazarus et al. 2009); I therefore rule out that this mechanism is at work here.

Although the lack of directional change and the lack of correspondence among the silicification trajectories point toward macroevolutionary processes above the lineage level, there are several complicating aspects to the results. First, two of the three lineages show net changes in silicification over time, in spite of showing weak support for the directional model, in both cases matching the direction of change in radiolarians overall. Could the assemblage-level decline result from a large number of lineages showing net changes toward less silicification, but displaying random-walk-like trajectory? This ought not to happen, statistically, since the random walk model that best fits the data from these lineages has a mean step size of zero; therefore, in a sufficiently large number of instantiations of the model, net changes ought to sum to zero also. However, if the model being fit were incorrect in that regard, net changes across many lineages could still add up to declining silicification of the whole assemblage.

A further complication is suggested by the remarkably close correspondence between the changes in silicification in the *Stichocorys* lineage and those in the whole radiolarian assemblage (Fig. 5.18, red squares and black circles, respectively). This correspondence suggests the opposite conclusion, namely that changes at the lineage level and at the assemblage level are linked—at least in some cases. Considering the role of abundance may help to explain this apparent contradiction: measurements of silicification in the whole-assemblage study (Lazarus et al. 2009) were made on a taxon-agnostic, per-individual basis, meaning that changes in assemblage silicification could also simply be explained by changes in the relative abundances of more and less silicified taxa (without invoking any evolutionary changes at all). Although this is certainly not the answer to the riddle, because the taxonomic composition of Paleocene radiolarian assemblages is quite different from those in the Pleistocene, it may help explain why some lineages track the whole-assemblage pattern closely and

others do not. This may simply be a reflection of the relative share of the whole assemblage those lineages constitute. Indeed, *Stichocorys* is a highly abundant taxon throughout much of its range, while *Centrobotrys*, which does not track the whole-assemblage pattern, is rare.

Examining the correspondence between the whole-assemblage silicification changes and those within the lineages examined highlights a final complicating factor in determining the causes behind the macroevolutionary trend: most of the change in the whole-assemblage silicification occurs from the mid-Eocene to the Oligocene. The subsequent time interval containing the silicification record of the individual lineages shows a much less obvious decline in whole-assemblage silicification, suggesting that within-lineage stasis or random walks in this particular part of the time-series may actually be compatible with an anagenetically-driven macroevolutionary pattern. If the decline in dissolved silica concentrations believed to drive this pattern (Harper and Knoll 1975) was confined to the Eocene/Oligocene transition (also a major oceanographic and climatic event, Zachos et al. 2001), we might not in fact expect to find directional evolution toward more lightly silicified tests in the Neogene.

Two lines of evidence from the fossil record of diatoms support a selection pressure on radiolarians that expanded in the Paleogene and subsequently stabilized. Reconstructions of diatom taxonomic diversity based on subsampling show a peak near the Eocene/Oligocene boundary (Rabosky and Sorhannus 2009 and Fig. 4.1 in Chapter 4). The reconstructions of diatom morphospace undertaken in Chapters 2 and 3 show less increase in the volume occupied in the Neogene as compared to the Paleogene (Fig. 2.9), particularly when sampling is taken into account (Fig. 3.4).

Further evidence suggesting high silica availability in the early Eocene, consistent with a later Eocene rise in diatom productivity and selection pressure on radiolaria, comes from an analysis of Cenozoic cherts by Muttoni and Kent (2007). The analysis suggests that inorganic, chemical chert precipitation occurred during the Early Eocene Climatic Optimum, resulting from elevated silica inputs to the ocean (derived from elevated weathering rates) and depressed

biological outputs (due to decreased upwelling and siliceous productivity). The decline in abiotic chert formation after a peak at ~50 Ma, accompanied by a peak in $\delta^{18}O$ values (and, by implication, peak temperature), coincides with the highest rates of decline in assemblage-level radiolarian silicification (Fig. 5.18).

In the course of selecting lineages for this study (described in Sections 5.2.3 and 5.3.1) I sought lineages spanning the interval of greatest change in silicification, but found very few. This dearth of described lineages crossing the Eocene/Oligocene boundary may, in part, be indicative of a turnover event. Explorations of the *Neptune* database spanning the Cenozoic found only a slight increase in origination rates near the boundary, but no discernible rise in extinction rate (Spencer-Cervato 1999, Figs. 4.13 and 4.14). More detailed studies, however, have reported turnover events both in the Antarctic (Lazarus et al. 2008), in the late Eocene, and at the Eocene/Oligocene boundary in the tropical Pacific (Funakawa et al. 2006). A turnover event around this time would seem to favor species selection and could explain the steep decline in silicification if extinction were biased toward heavily silicified radiolarians, and originations biased in the opposite direction. This hypothesis warrants further attention, but would require significant taxonomic and biostratigraphic work in addition to establishing silicification by morphometrics, in the manner carried out here, for a significant proportion of the Eocene and Oligocene assemblages.

5.6 Conclusions

1. Changes in silicification in the three radiolarian lineages examined here do not strongly support the model for directional evolution; two lineages support the random walk model, while one supports the model for stasis. These results suggest that macroevolutionary processes above the species level are responsible for assemblage-wide decrease in silicification.

2. The different patterns among lineages may be attributable to biological differences in the role of the test in feeding ecology.

3. No relationship was found between pore area and thickness, unlike in diatoms, suggesting that the radiolarian test plays a different biological role than the diatom frustule.

4. It is unlikely that diffusion from a bounded origin is the macroevolutionary process underlying the assemblage-level decline in silicification, because maximum silicification decreases in concert with mean silicification.

5. Although these results point toward selection among, and not within, lineages, three caveats caution against ruling out a role for anagenesis:

   (a) Two of the lineages do show net changes. If the model fit to these data is inaccurate, such net silicification changes in many lineages could still add up to overall decrease in silicification.

   (b) Variations in species abundance may play a role in explaining the whole-assemblage pattern. For instance, *Stichocorys* is very abundant and also closely tracks the whole-assemblage trajectory in silicification.

   (c) Most of the change in the assemblage-level trajectory occurs in the Paleogene, especially around the Eocene/Oligocene boundary. It does not show much change in the Neogene, where the lineages examined here occur; we thus might not expect to see much directional change in these lineages even if the assemblage-level trend was the result of within-species selection.

6. There is some evidence for a turnover event at the Eocene/Oligocene boundary, which could explain the assemblage-level pattern if there was biased extinction of highly silicified lineages and biased origination of lightly silicified lineages. Further work is needed to quantify silicification in many lineages on either side of this boundary, once the requisite taxonomic and phylogenetic relationships have been established.

References

Abràmoff, M. D., P. J. Magalhães, and S. J. Ram. 2004. Image processing with ImageJ. Biophotonics International 11:36–42.

Akiba, F. 1986. Taxonomy, morphology and phylogeny of the Neogene diatom zonal marker species in the middle-to-high latitudes of the North Pacific. Initial Reports of the Deep Sea Drilling Project 87:483–554.

Alroy, J. 1996. Constant extinction, constrained diversification, and uncoordinated stasis in North American mammals. Palaeogeography, Palaeoclimatology, Palaeoecology 127:285–311.

———. 1998. Cope's rule and the dynamics of body mass evolution in North American fossil mammals. Science 280:731–734.

———. 2000. New methods for quantifying macroevolutionary patterns and processes. Paleobiology 26:707–799.

———. 2010a. Fair sampling of taxonomic richness and unbiased estimation of origination and extinction rates. *In* J. Alroy and G. Hunt, eds., Quantitative Methods in Paleobiology, Paleontological Society Papers, 16:55–80.

———. 2010b. Geographical, environmental and intrinsic biotic controls on Phanerozoic marine diversification. Palaeontology 53:1211–1235.

———. 2010c. The shifting balance of diversity among major marine animal groups. Science 329:1191–1194.

Alroy, J., C. R. Marshall, R. K. Bambach, K. Bezusko, M. Foote, F. T. Fürsich, T. A. Hansen, S. M. Holland, L. C. Ivany, D. Jablonski, et al. 2001. Effects of sampling standardization on estimates of Phanerozoic marine diversification. Proceedings of the National Academy of Sciences 98:6261.

Alroy, J., M. Aberhan, D. Bottjer, M. Foote, F. Fürsich, P. Harries, A. Hendy, S. Holland, L. Ivany, W. Kiessling, et al. 2008. Phanerozoic trends in the global diversity of marine invertebrates. Science 321:97–100.

Alverson, A. J. 2007. Strong purifying selection in the silicon transporters of marine and freshwater diatoms. Limnology and Oceanography 1420–1429.

Anderson, B. M., D. Pisani, A. I. Miller, and K. J. Peterson. 2011. The environmental affinities of marine higher taxa and possible biases in their first appearances in the fossil record. Geology 39:971–974.

Anderson, O. 1983. Radiolaria. Springer Verlag, New York.

Andrews, G. W. 1976. Miocene marine diatoms from the Choptank Formation, Calvert County, Maryland. U.S. Geological Survey Professional Paper 910.

———. 1988. A revised marine diatom zonation for Miocene strata of the southeastern United States. U.S. Geological Survey Professional Paper 1481.

Anonymous. 1962. Report of the Systematics Association committee for descriptive biological terminology, II and IIa. Terminology of simple symmetrical plane shapes (Charts 1, 1a). Taxon 11:145–156.

———. 1975. Proposals for a standardization of diatom terminology and diagnoses. Beihefte zur Nova Hedwigia 53:323–354.

Arita, S. and T. Ohtsuka. 2004. Describing the valve outlines of *Navicula* species using a newly described arc-constitutive model. Diatom 20:191–198.

Arnold, A. J., D. C. Kelly, and W. C. Parker. 1995. Causality and Cope's rule: evidence from the planktonic foraminifera. Journal of Paleontology 203–210.

Baldauf, J. G. and J. A. Barron. 1987. Oligocene marine diatoms recovered in dredge samples from the Navarin Basin Province, Bering Sea. USGS Bulletin 1765.

Bambach, R. 1999. Energetics in the global marine fauna: A connection between terrestrial diversification and change in the marine biosphere. Geobios 32:131–144.

Bambach, R. K., A. H. Knoll, and J. J. Sepkoski. 2002. Anatomical and ecological constraints on Phanerozoic animal diversity in the marine realm. Proceedings of the National Academy of Sciences 99:6854–6859.

Barber, H. G. and E. Y. Haworth. 1981. A guide to the morphology of the diatom frustule: with a key to the British freshwater genera. Freshwater Biological Association.

Barron, J. and A. Mahood. 1993. Exceptionally well-preserved early Oligocene diatoms from glacial sediments of Prydz Bay, East Antarctica. Micropaleontology 29–45.

Barron, J. A. 1976. Revised Miocene and Pliocene diatom biostratigraphy of upper Newport Bay, Newport Beach, California. Marine Micropaleontology 1:27–63.

———. 1981. Late Cenozoic diatom biostratigraphy and paleoceanography of the middle-latitude eastern North Pacific, Deep Sea Drilling Project Leg 63. Initial Reports of the Deep Sea Drilling Project 63:507–538.

———. 1985. Miocene to Holocene planktic diatoms. *In* H. Bolli, J. Saunders, and K. Perch-Nielsen, eds., Plankton Stratigraphy, 713–762. Cambridge University Press, Cambridge, U.K.

Berggren, W. A., D. V. Kent, C. C. Swisher, and M.-P. Aubry. 1995. A revised Cenozoic geochronology and chronostratigraphy. Geochronology, Time Scales, and Global Stratigraphic Correlation 129–212.

Bolli, H. M., J. B. Saunders, and K. Perch-Nielsen. 1989. Plankton stratigraphy: Planktic foraminifera, calcareous nannofossils and calpionellids, vol. 1. Cambridge University Press, Cambridge, U.K.

Boyce, C. K. and A. H. Knoll. 2002. Evolution of developmental potential and the multiple independent origins of leaves in Paleozoic vascular plants. Paleobiology 28:70.

Burckle, L. H. 1974. Size differences in the diatom, *Annellus californicus* Tempere. Trans. Proc. Palaeont. Soc. Japan 96:437–441.

Bush, A., M. Markey, and C. Marshall. 2004. Removing bias from diversity curves: the effects of spatially organized biodiversity on sampling-standardization. Paleobiology 30:666–686.

Butler, R., S. Brusatte, B. Andres, and R. Benson. 2012. How do geological sampling biases affect studies of morphological evolution in deep time? A case study of pterosaur (Reptilia: Archosauria) disparity. Evolution 66:147–162.

Chacón-Baca, E., H. Beraldi-Campesi, S. R. S. Cevallos-Ferriz, A. H. Knoll, and S. Golubic. 2002. 70 Ma nonmarine diatoms from northern Mexico. Geology 30:279.

Chang, K. H., K. Suzuki, S. O. Park, K. Ishida, and K. Uno. 2003. Recent advances in the Cretaceous stratigraphy of Korea. Journal of Asian Earth Sciences 21:937–948.

Ciampaglio, C. N., M. Kemp, and D. W. McShea. 2001. Detecting changes in morphospace occupation patterns in the fossil record: characterization and analysis of measures of disparity. Paleobiology 27:695–715.

Cox, E. J. 2006. *Achnanthes* sensu stricto belongs with genera of the Mastogloiales rather than with other monoraphid diatoms (Bacillariophyta). European Journal of Phycology 41:67–81.

Damsté, J. S. S., G. Muyzer, B. Abbas, S. W. Rampen, G. Massé, W. G. Allard, S. T. Belt, J. M. Robert, S. J. Rowland, J. M. Moldowan, et al. 2004. The rise of the rhizosolenid diatoms. Science 304:584–587.

Darwin, C. 1859. On the Origin of the Species by Means of Natural Selection: Or, The Preservation of Favoured Races in the Struggle for Life. John Murray, London.

De Stefano, M., W. H. C. F. Kooistra, and D. Marino. 2003. Morphology of the diatom genus *Campyloneis* (Bacillariophyceae, Bacillariophyta), with a description of *Campyloneis juliae* sp. nov. and an evaluation of the function of the valvocopulae. Journal of Phycology 39:735–753.

Deline, B. 2009. The effects of rarity and abundance distributions on measurements of local morphological disparity. Paleobiology 35:175–189.

DeMaster, D. 2003. The diagenesis of biogenic silica: chemical transformations occurring in the water column, seabed, and crust. *In* H. D. Holland and K. K. Turekian, eds., Treatise on Geochemistry, vol. 7, 87–98. Elsevier.

Drebes, G. and D. Schulz. 1989. *Anaulus australis* sp. nov. (Centrales, Bacillariophyceae), a new marine surf zone diatom, previously assigned to *A. birostratus* (Grunow) Grunow. Botanica marina 32:53–64.

Du Buf, H. and M. M. Bayer. 2002. Automatic Diatom Identification, *Series in Machine Perception and Artificial Intelligence*, vol. 51. World Scientific, Singapore.

Dudley, R. and C. Gans. 1991. A critique of symmorphosis and optimality models in physiology. Physiological Zoology 64:627–637.

Dugdale, R. C. and F. P. Wilkerson. 1998. Silicate regulation of new production in the equatorial Pacific upwelling. Nature 391:270–273.

Eble, G. 2000a. Contrasting evolutionary flexibility in sister groups: disparity and diversity in Mesozoic atelostomate echinoids. Paleobiology 26:56–079.

———. 2000b. Theoretical morphology: state of the art. Paleobiology 26:520–528.

Edelsbrunner, H. and E. P. Mücke. 1992. Three-dimensional alpha shapes. *In* Proceedings of the 1992 Workshop on Volume Visualization, 75–82. ACM.

Ehrenberg, C. G. 1838. Die Infusionsthierchen als vollkommene Organismen. Ein Blick in das tiefere organische Leben der Natur. Leopold Voss, Leipzig.

Eldredge, N. and S. Gould. 1972. Punctuated equilibria: an alternative to phyletic gradualism. *In* T. Schopf, ed., Models in Paleobiology, 82–115. Freeman, Cooper & Co.

Erwin, D. 2007. Disparity: morphological pattern and developmental context. Palaeontology 50:57–73.

Erwin, D. H. 2000. Macroevolution is more than repeated rounds of microevolution. Evolution & Development 2:78–84.

Erwin, D. H., M. Laflamme, S. M. Tweedt, E. A. Sperling, D. Pisani, and K. J. Peterson. 2011. The Cambrian conundrum: Early divergence and later ecological success in the early history of animals. Science 334:1091–1097.

Falkowski, P. G., M. E. Katz, A. H. Knoll, A. Quigg, J. A. Raven, O. Schofield, and F. J. R. Taylor. 2004. The Evolution of Modern Eukaryotic Phytoplankton. Science 305:354–360.

Fenner, J. 1984. Eocene-Oligocene planktic diatom stratigraphy in the low latitudes and the high southern latitudes. Micropaleontology 319–342.

———. 1985. Late Cretaceous to Oligocene planktic diatoms. *In* H. Bolli, J. Saunders, and K. Perch-Nielsen, eds., Plankton Stratigraphy, 713–762. Cambridge University Press, Cambridge, U.K.

———. 1991a. Taxonomy, stratigraphy, and paleoceanographic implications of paleocene diatoms. Initial Reports of the Deep Sea Drilling Project 114:123–154.

Fenner, J. M. 1991b. Late Pliocene-Quaternary quantitative diatom stratigraphy in the Atlantic sector of the Southern Ocean. Proceedings of the Ocean Drilling Program, Scientific Results 114:97–121.

Finkel, Z. and B. Kotrc. 2010. Silica use through time: macroevolutionary change in the morphology of the diatom frustule. Geomicrobiology Journal 27:596–608.

Finkel, Z. V., M. E. Katz, J. D. Wright, O. M. Schofield, and P. G. Falkowski. 2005. Climatically driven macroevolutionary patterns in the size of marine diatoms over the Cenozoic. Proceedings of the National Academy of Sciences of the United States of America 102:8927–8932.

Flessa, K. and D. Jablonski. 1983. Extinction is here to stay. Paleobiology 9:315–321.

Follows, M., S. Dutkiewicz, S. Grant, and S. Chisholm. 2007. Emergent biogeography of microbial communities in a model ocean. Science 315:1843–1846.

Foote, M. 1989. Perimeter-based Fourier analysis: a new morphometric method applied to the trilobite cranidium. Journal of Paleontology 880–885.

———. 1992. Rarefaction analysis of morphological and taxonomic diversity. Paleobiology 18:1–16.

———. 1993. Discordance and concordance between morphological and taxonomic diversity. Paleobiology 19:185–204.

———. 1995a. Morphological diversification of Paleozoic crinoids. Paleobiology 21:273–299.

———. 1995b. Morphology of Carboniferous and Permian crinoids. Contributions from the Museum of Paleontology, University of Michigan 29:135–184.

———. 1997. The evolution of morphological diversity. Annual Review of Ecology and Systematics 28:129–152.

———. 1999. Morphological diversity in the evolutionary radiation of Paleozoic and post-Paleozoic crinoids. Paleobiology 25:1–116.

Foreman, H. P. 1973. Radiolaria of Leg 10 with systematics and ranges for the families Amphipyndacidae, Artostrobiidae, and Theoperidae. Initial Reports of the Deep Sea Drilling Project 10:407–474.

Fourtanier, E. 1991. Diatom biostratigraphy of equatorial Indian Ocean Site 758. Ocean Drilling Program Scientific Results 121:189–208.

Funakawa, S., H. Nishi, T. C. Moore, and C. A. Nigrini. 2006. Radiolarian faunal turnover and paleoceanographic change around Eocene/Oligocene boundary in the central equatorial Pacific, ODP Leg 199, Holes 1218A, 1219A, and 1220A. Palaeogeography, Palaeoclimatology, Palaeoecology 230:183–203.

Garcia, M. 2004. Morphology and taxonomy of *Neohuttonia reichardtii* (Grunow) O. Kuntze (Bacillariophyta) from southern Brazil. Iheringia, Série Botânica 59:179–182.

Gersonde, R. and D. M. Harwood. 1990. Lower Cretaceous diatoms from ODP Leg 113 Site 693 (Weddell Sea). Part 1: Vegetative cells. Proceedings of the Ocean Drilling Program, Scientific Results 113:365–402.

Gingerich, P. D. 1989. New earliest Wasatchian mammalian fauna from the Eocene of northwestern Wyoming: composition and diversity in a rarely sampled high-floodplain assemblage high-floodplain assemblage. University of Michigan Papers on Paleontology 28:1–97.

Gombos, A. M. 1980. The early history of the diatom family Asterolampraceae. Bacillaria 3:227–272.

Gombos Jr, A. M. and P. F. Ciesielski. 1983. Late Eocene to early Miocene diatoms from the southwest Atlantic. Initial Reports of the Deep Sea Drilling Project 71:583–634.

Good, I. 1953. The population frequencies of species and the estimation of population parameters. Biometrika 40:237–264.

Gould, S. 1992. Bully for Brontosaurus: Reflections in Natural History. W W Norton & Company Incorporated.

Gould, S. J. 1988. Trends as changes in variance: a new slant on progress and directionality in evolution. Journal of Paleontology 319–329.

———. 1989. Wonderful Life: The Burgess Shale and the Nature of History. W.W. Norton.

Gould, S. J. and N. Eldredge. 1993. Punctuated equilibrium comes of age. Nature 366:223–227.

Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325–338.

Greville, R. K. 1863. Descriptions of new and rare diatoms. Series VIII. Transactions of the Microscopical Society & Journal 11:13–21.

Guiry, M. D. and G. M. Guiry. 2011. AlgaeBase. World-wide electronic publication, National University of Ireland, Galway, http://www.algaebase.org.

Haeckel, E. 1887. Report on the Radiolaria collected by HMS Challenger during the years 1873–1876. *In* C. W. Thompson and J. Murray, eds., Report on the Scientific Results of the Voyage of the HMS Challenger, Zoology, 18, 2 parts. Her Majesty's Stationery Office, London.

———. 1904. Kunstformen der Natur. Bibliographisches Institut, Leipzig/Vienna.

Hajós, M. 1976. Upper Eocene and Lower Oligocene Diatomaceae, Archaeomodaceae, and Silicoflagellatae in Southwestern Pacific sediments. Initial Reports of the Deep Sea Drilling Project 35:817–883.

Hajós, M. and H. Stradner. 1975. Late Cretaceous Archaeomonadaceae, Diatomaceae, and Silicoflagellatae from the South Pacific Ocean, Deep Sea Drilling Project, Leg 29, Site 275. Initial Reports of the Deep Sea Drilling Project 29:913–1009.

Hallam, A. 1978. How rare is phyletic gradualism and what is its evolutionary significance? Evidence from Jurassic bivalves. Paleobiology 16–25.

Hamm, C. and V. Smetacek. 2007. Armor: Why, When, and How. *In* P. G. Falkowski and A. H. Knoll, eds., Evolution of Primary Producers in the Sea. Elsevier, Burlington, MA.

Hamm, C., R. Merkel, O. Springer, P. Jurkojc, C. Maier, K. Prechtel, and V. Smetacek. 2003. Architecture and material properties of diatom shells provide effective mechanical protection. Nature 421:841–843.

Hanna, G. D. 1927. Cretaceous diatoms from California. Occasional Papers of the California Academy of Sciences 13:5–40.

———. 1930. A review of the genus *Rouxia*. Journal of Paleontology 4:179–188.

———. 1932. The diatoms of Sharktooth Hill, Kern County, California. Proceedings of the California Academy of Sciences 4:161–263.

Hansen, T. A. 1982. Modes of larval development in early Tertiary neogastropods. Paleobiology 367–377.

Hargraves, P. E. 1986. The relationship of some fossil diatom genera to resting spores. *In* M. Ricard, ed., Proc. 8th Int. Diatom Symp., Paris, Aug. 1984, 27:67–80. Koeltz Scientific, Königstein, Germany.

Harper, H. E. and A. H. Knoll. 1975. Silica, diatoms, and Cenozoic radiolarian evolution. Geology 3:175–177.

Harwood, D. M. 1988. Upper Cretaceous and lower Paleocene diatom and silicoflagellate biostratigraphy of Seymour Island, eastern Antarctic Peninsula. Geological Society of America.

Harwood, D. M., V. A. Nikolaev, and D. M. Winter. 2007. Cretaceous records of diatom evolution, radiation, and expansion. *In* S. W. Starratt, ed., Pond scum to carbon sink: Geological and environmental applications of the diatoms, Paleontological Society Papers, 13:33–59. The Paleontological Society.

Hasle, G. R. 1975. Some living marine species of the diatom family Rhizosoleniaceae. Beihefte zur Nova Hedwigia 53:99–140.

Hasle, G. R. and P. A. Sims. 1986. The diatom genera *Stellarima* and *Symbolophora* with comments on the genus *Actinoptychus*. British Phycological Journal 21:97–114.

Hendey, N. I. 1958. Marine diatoms from some West African ports. Journal of the Royal Microscopical Society (Great Britain) 77:28.

———. 1969. *Pyrgupyxis*: A new genus of diatoms from a South Atlantic Eocene Core. Occasional Papers of the California Academy of Sciences 72.

———. 1981. Note on the genus *Neobrunia* O. Kuntze. Bacillaria 4:7–20.

Hendey, N. I. and R. Simonsen. 1972. *Muelleriella limbata* (Ehrenberg) Van Heurck in Eocene South Atlantic Cores. Nova Hedwigia 39:79–94.

Hunt, G. 2006. Fitting and comparing models of phyletic evolution: random walks and beyond. Paleobiology 32:578–601.

———. 2007. The relative importance of directional change, random walks, and stasis in the evolution of fossil lineages. Proceedings of the National Academy of Sciences 104:18404–18408.

———. 2008. Gradual or pulsed evolution: when should punctuational explanations be preferred? Paleobiology 34:360–377.

———. 2012. paleoTS: Analyze paleontological time-series. R package version 0.4-4.

Huntley, J. W., S. Xiao, and M. Kowalewski. 2006. 1.3 billion years of acritarch history: An empirical morphospace approach. Precambrian Research 144:52–68.

Hustedt, F. and N. G. Jensen. 1985. The pennate diatoms: a translation of Hustedt's "Die Kieselalgen, 2. Teil", vol. 2. Koeltz Scientific Books.

Hutchinson, G. 1978. An introduction to population ecology. Yale University Press New Haven.

Jablonski, D. 1997. Body-size evolution in Cretaceous molluscs and the status of Cope's rule. Nature 385:250–252.

———. 2007. Scale and hierarchy in macroevolution. Palaeontology 50:87–109.

Jacot Des Combes, H. and A. Abelmann. 2009. From species abundance to opal input: Simple geometrical models of radiolarian skeletons from the atlantic sector of the southern ocean. Deep Sea Research Part I: Oceanographic Research Papers 56:757–771.

Jahn, R. and W. H. Kusber. 2005. Reinstatement of the genus *Ceratoneis* Ehrenberg and lectotypification of its type specimen: *C. closterium* Ehrenberg. Diatom Research 20:295–304.

Jousé, A. P. 1974. Diatoms in the Oligocene-Miocene biostratigraphic zones of the tropical areas of the Pacific Ocean. Beihefte zur Nova Hedwigia 45:333–364.

Kling, S. 1978. Radiolaria. *In* B. Haq and A. Boersma, eds., Introduction to Marine Micropaleontology, 203–244. Elsevier Science.

Knoll, A. 2003. Biomineralization and Evolutionary History. Reviews in Mineralogy and Geochemistry 54:329–356.

Knoll, A. H. and D. A. Johnson. 1975. Late Pleistocene evolution of the collosphaerid radiolarian *Buccinosphaera invaginata* Haeckel. Micropaleontology 60–68.

Knoll, A. H., R. Bambach, D. Canfield, J. Grotzinger, et al. 1996. Comparative Earth history and Late Permian mass extinction. Science 452–457.

Koizumi, I. 1973. The late Cenozoic diatoms of Sites 183-193, Leg 19, Deep Sea Drilling Project. Initial Reports of the Deep Sea Drilling Project 19:805–855.

Komura, S. 1975. *Ikebea*, eine neue Gattung der pennaten Bacillariaceen aus dem Neogen Japans. Transactions and Proceedings of the Palaeontological Society of Japan 99:133–142.

———. 1976. *Sawamuraia, Katahiraia* und *Yoshidaia*, drei neue Diatomgattungen aus dem Neogen Japans. Transactions and Proceedings of the Palaeontological Society of Japan 103.

Kooistra, W., R. Gersonde, L. K. Medlin, and D. G. Mann. 2007. The origin and evolution of the diatoms: their adaptation to a planktonic existence. *In* P. G. Falkowski and A. H. Knoll, eds., Evolution of Primary Producers in the Sea. Elsevier, Boston.

Kooistra, W. H. C. F. and L. Medlin. 1996. Evolution of the diatoms (Bacillariophyta) IV: A reconstruction of their age from small subunit rRNA coding regions and fossil record. Molecular Phylogenetics and evolution 6:391–407.

Lafarge, T. and B. Pateiro-Lopez. 2012. alphashape3d: Implementation of the 3D alpha-shape for the reconstruction of 3D sets from a point cloud. R package version 1.0.

Lazarus, D. 2011. The deep-sea microfossil record of macroevolutionary change in plankton and its study. Geological Society, London, Special Publications 358:141–166.

Lazarus, D., C. Hollis, and M. Apel. 2008. Patterns of opal and radiolarian change in the Antarctic mid-Paleogene: Clues to the origin of the Southern Ocean. Micropaleontology 41–48.

Lazarus, D., J. A. Barron, A. Türke, P. Diver, and J. Renaudie. 2012a. Diversity history of cenozoic planktic marine diatoms. *In* The Micropalaeontological Society AGM and Warm World Symposium, British Geological Survey, Nottingham, UK, Nov. 11th-13th. The Micropalaeontological Society.

Lazarus, D. B. 1986. Tempo and mode of morphologic evolution near the origin of the radiolarian lineage Pterocanium prismatium. Paleobiology 175–189.

———. 1994. Neptune: a marine micropaleontology database. Mathematical Geology 26:817–832.

———. 2005. A brief review of radiolarian research. Paläontologische Zeitschrift 79:183–200.

Lazarus, D. B., B. Kotrc, G. Wulf, and D. N. Schmidt. 2009. Radiolarians decreased silicification as an evolutionary response to reduced Cenozoic ocean silica availability. Proceedings of the National Academy of Sciences 106:9333.

Lazarus, D. B., M. Weinkauf, and P. Diver. 2012b. Pacman profiling: a simple procedure to identify stratigraphic outliers in high-density deep-sea microfossil data. Paleobiology 38:144–161.

Lewontin, R. 1969. The meaning of stability. *In* Brookhaven Symposia in Biology, 22:13.

Liow, L. and N. Stenseth. 2007. The rise and fall of species: implications for macroevolutionary and macroecological studies. Proceedings of the Royal Society B: Biological Sciences 274:2745–2752.

Liow, L., H. Skaug, T. Ergon, and T. Schweder. 2010. Global occurrence trajectories of microfossils: environmental volatility and the rise and fall of individual species. Paleobiology 36:224–252.

Llano, G. A. and I. E. Wallen. 1971. Biology of the Antarctic Seas IV, vol. 4. American Geophysical Union.

Lloyd, G., P. Pearson, J. Young, and A. Smith. 2012a. Sampling bias and the fossil record of planktonic foraminifera on land and in the deep sea. Paleobiology 38:569–584.

Lloyd, G., J. Young, and A. Smith. 2012b. Comparative quality and fidelity of deep-sea and land-based nannofossil records. Geology 40:155–158.

Low, S. L. 2006. Quantifying the morphological evolution of the Nautiloidea through the Phanerozoic. Ph.D. thesis, Harvard University.

Lupia, R. 1999. Discordant morphological disparity and taxonomic diversity during the Cretaceous angiosperm radiation: North American pollen record. Paleobiology 25:1–28.

MacFadden, B. J. 1985. Patterns of phylogeny and rates of evolution in fossil horses: hipparions from the Miocene and Pliocene of North America. Paleobiology 245–257.

———. 1994. Fossil horses: systematics, paleobiology, and evolution of the family Equidae. Cambridge University Press.

Mann, D. and S. Droop. 1996. Biodiversity, biogeography and conservation of diatoms. Hydrobiologia 336:19–32.

Marshall, C. R. 1990. Confidence intervals on stratigraphic ranges. Paleobiology 1–10.

———. 2003. Nomothetism and understanding the Cambrian "explosion". Palaios 18:195–196.

Marx, F. G. and M. D. Uhen. 2010. Climate, critters, and cetaceans: Cenozoic drivers of the evolution of modern whales. Science 327:993–996.

Matsuoka, A. 2007. Living radiolarian feeding mechanisms: new light on past marine ecosystems. Swiss Journal of Geoscience 100:273–279.

McGhee, G. R. 1999. Theoretical morphology: the concept and its applications. Columbia University Press.

McShea, D. W. 1994. Mechanisms of large-scale evolutionary trends. Evolution 1747–1763.

Medlin, L. K. and I. Kaczmarska. 2004. Evolution of the diatoms: V. Morphological and cytological support for the major clades and a taxonomic revision. Phycologia 43:245–270.

Meyer, D., A. Zeileis, and K. Hornik. 2011. vcd: Visualizing Categorical Data. R package version 1.2-12.

Miller, A. and M. Foote. 1996. Calibrating the Ordovician radiation of marine life: implications for Phanerozoic diversity trends. Paleobiology 304–309.

Moore, T. 1969. Radiolaria; change in skeletal weight and resistance to solution. Bulletin of the Geological Society of America 80:2103–2107.

———. 1971. Radiolaria. Initial Reports of the Deep Sea Drilling Project 8:727–775.

Moshkovitz, S., A. Ehrlich, and D. Soudry. 1983. Siliceous microfossils of the Upper Cretaceous Mishash Formation, Central Negev, Israel. Cretaceous Research 4:173–194.

Mou, D. and E. Stoermer. 2004. Separating *Tabellaria* (Bacillariophyceae) shape groups based on Fourier descriptors. Journal of Phycology 28:386–395.

Muttoni, G. and D. V. Kent. 2007. Widespread formation of cherts during the early Eocene climate optimum. Palaeogeography, Palaeoclimatology, Palaeoecology 253:348–362.

Netherlands Architecture Institute. 2012. Crown for Electric Light by Hendricus Petrus Berlage.
`http://schatkamer.nai.nl/system/pictures/546/original/BERL_251-5_900px.jpg?1348145419`.

Newell, N. 1959. Adequacy of the fossil record. Journal of Paleontology 488–499.

Niklas, K. 1999. Evolutionary walks through a land plant morphospace. Journal of Experimental Botany 50:39–52.

Niklas, K. J. 2004. Computer models of early land plant evolution. Annual Review of Earth and Planetary Sciences 32:47–66.

Norris, R. D. 1991. Biased extinction and evolutionary trends. Paleobiology 388–399.

Olney, M. P., R. P. Scherer, S. M. Bohaty, and D. M. Harwood. 2005. Eocene-Oligocene paleoecology and the diatom genus Kisseleviella Sheshukova-Poretskaya from the Victoria Land Basin, Antarctica. Marine Micropaleontology 58:56–72.

Olney, M. P., R. P. Scherer, D. M. Harwood, and S. M. Bohaty. 2007. Oligocene–early Miocene Antarctic nearshore diatom biostratigraphy. Deep Sea Research Part II: Topical Studies in Oceanography 54:2325–2349.

Olshtynskaja, A. P. and H. Simola. 1990. Morphology of the diatom genus *Pseudopodosira*. *In* Proceedings of the 10th International Diatom Symposium, Joensuu, Finland, August 28-September 2, 1998, 93. Balogh Scientific Books.

Olshtynskaya, A. P. 2002. Morphological and taxonomic characteristics of some Paleogene diatoms of Ukraine. International Journal on Algae 4:118–126.

Pantocsek, J. 1886. Beiträge Zur Kenntniss Der Fossilen Bacillarien Ungarns, vol. 1. Buchdruckerei von Julius Platzko.

Pappas, J. L. 2005. Theoretical morphospace and its relation to freshwater Gomphonemoid–Cymbelloid diatom (Bacillariophyta) lineages. Journal of Biological Systems 13:385–398.

Patrick, R. and C. W. Reimer. 1975. The Diatoms of the United States II, vol. 1. Acad. Nat. Sci. Philad. Monogr.

Peters, S. E. 2004. Relative abundance of Sepkoski's evolutionary faunas in Cambrian–Ordovician deep subtidal environments in North America. Paleobiology 30:543–600.

Petrushevskaya, M. 1965. Peculiarities of the construction of the skeleton of radiolarians Botryoidae (Order Nassellaria). Transactions of the Institute of Zoology, Academy of Sciences, USSR (NAUKA) 35:79–118.

Proctor, R. 2006. Architecture from the cell-soul: Renè Binet and Ernst Haeckel. The Journal of Architecture 11:407–424.

Prothero, D. and N. Shubin. 1989. The evolution of Oligocene horses. *In* D. Prothero and R. Schoch, eds., The Evolution of Perissodactyls, 142–175. Oxford University Press.

R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rabosky, D. L. and U. Sorhannus. 2009. Diversity dynamics of marine planktonic diatoms across the Cenozoic. Nature 457:183–186.

Raup, D. M. 1972. Taxonomic diversity during the Phanerozoic. Science 177:1065–1071.

Raup, D. M. and S. J. Gould. 1974. Stochastic simulation and evolution of morphology—towards a nomothetic paleontology. Systematic Biology 23:305–322.

Raup, D. M. and A. Michelson. 1965. Theoretical morphology of the coiled shell. Science 147:1294–1295.

Raven, J. A. 1997. The vacuole: A cost-benefit analysis. Advances in Botanical Research 25:59–86.

Raven, J. A. and A. M. Waite. 2004. The evolution of silicification in diatoms: inescapable sinking and sinking as escape? New Phytologist 162:45–61.

Riedel, W. 1967. Some new families of Radiolaria. *In* Proceedings of the Geological Society of London, 1640:148–149.

Riedel, W. and A. Sanfilippo. 1971. Cenozoic Radiolaria from the western tropical Pacific, Leg 7. Initial Reports of the Deep Sea Drilling Project 7:1529–1672.

———. 1978. Stratigraphy and evolution of tropical Cenozoic radiolarians. Micropaleontology 61–96.

Rohlf, F. and F. Bookstein. 1990. Proceedings of the Michigan Morphometrics Workshop. University of Michigan Museum of Zoology.

Ross, R. and P. A. Sims. 1980. *Dextradonator* Ross & Sims, nov. gen. and *Abas* Ross & Sims, nov. gen. Bacillaria 3:115–127.

———. 1987. Further genera of the Biddulphiaceae (diatoms) with interlocking linking spines. British Museum (Natural History).

Rothpletz, A. 1896. Ueber die Flysch-Fucoiden und einige andere fossile Algen, sowie über liasische Diatomeen führende Hornschwämme. Zeitschrift der Deutschen Geologischen Gesellschaft 48:854–914.

Round, F. E., R. M. Crawford, and D. G. Mann. 1990. The Diatoms: biology & morphology of the genera. Cambridge University Press.

Sabbe, K. and W. Vyverman. 1995. Taxonomy, morphology and ecology of some widespread representatives of the diatom genus *Opephora*. European Journal of Phycology 30:235–249.

Sanfilippo, A. 1990. Origin of the subgenera *Cyclampterium, Paralampterium* and *Sciadiopeplus* from *Lophocyrtis (Lophocyrtis)* (Radiolaria, Theoperidae). Marine Micropaleontology 15:287–312.

Sanfilippo, A. and W. Riedel. 1970. Post-Eocene "closed" theoperid radiolarians. Micropaleontology 446–462.

———. 1973. Cenozoic Radiolaria (exclusive of theoperids, artostrobiids and amphipyndacids) from the Gulf of Mexico, DSDP Leg 10. Initial Reports of the Deep Sea Drilling Project 10:475–611.

———. 1980. A revised generic and suprageneric classification of the Artiscins (Radiolaria). Journal of Paleontology 1008–1011.

———. 1992. The origin and evolution of Pterocorythidae (Radiolaria): A Cenozoic phylogenetic study. Micropaleontology 1–36.

Sanfilippo, A., M. Westberg-Smith, and W. Riedel. 1985. Cenozoic Radiolaria. *In* H. Bolli, J. Saunders, and K. Perch-Nielsen, eds., Plankton Stratigraphy, 631–712. Cambridge University Press.

Scherer, R. P. and N. Koç. 1996. Late Paleogene diatom biostratigraphy and paleoenvironments of the northern Norwegian-Greenland Sea. *In* Proceedings of the Ocean Drilling Program Scientific Results, 151:75–99. Ocean Drilling Program.

Schmid, A. M. M. 2007. The "paradox" diatom *Bacillaria paxillifer* (Bacillariophyta) revisited. Journal of Phycology 43:139–155.

Schmidt, D., H. Thierstein, J. Bollmann, and R. Schiebel. 2004. Abiotic forcing of plankton evolution in the cenozoic. Science 303:207–210.

Schrader, H. J. 1974. Cenozoic marine planktonic diatom stratigraphy of the tropical Indian Ocean. Initial Reports of the Deep Sea Drilling Project 24:887–968.

Schrader, H. J. and J. Fenner. 1976. Norwegian Sea Cenozoic diatom biostratigraphy and taxonomy. Initial Reports of the Deep Sea Drilling Project 38:989.

Sepkoski Jr, J. J. 1988. Alpha, beta, or gamma: where does all the diversity go? Paleobiology 221–234.

Shen, B., L. Dong, S. Xiao, and M. Kowalewski. 2008. The Avalon explosion: evolution of Ediacara morphospace. Science 319:81.

Sheshukova-Poretskaya, V. S. 1962. New and rare Bacillariophyta from the Diatom Suite of Sakhalin. [russian and latin.]. Uchenyye Zapiski, Seriya Biologicheskii Nauk 313:203–211.

Siegel, S. and N. J. Castellan Jr. 1988. Nonparametric Statistics for the Behavioral Sciences. McGraw Hill, 2 ed.

Simonsen, R. 1979. The diatom system: ideas on phylogeny. Bacillaria 2:9–71.

———. 1982. Note on the diatom genus *Charcotia* M. Peragallo. Bacillaria 5:101–116.

Sims, P. A. 1986. *Sphynctolethus* Hanna, *Ailuretta* gen. nov., and evolutionary trends within the Hemiauloideae. Diatom Research 1:241–269.

———. 1988. The fossil genus *Trochosira*, its morphology, taxonomy, and systematics. Diatom Research 3:245–257.

———. 1990. The fossil diatom genus *Fenestrella*, its morphology, systematics and palaeogeography. Beihefte zur Nova Hedwigia 100:277–288.

———. 2006. A revision of the genus *Rattrayella* De-Toni including a discussion on related genera. Diatom Research 21:125–158.

Sims, P. A., D. G. Mann, and L. K. Medlin. 2006. Evolution of the diatoms: insights from fossil, biological and molecular data. Phycologia 45:361–402.

Small, J. 1946. Quantitative evolution: Numerical analysis of tables to illustrate the geological history of species number in diatoms; an introductory summary. *In* Proceedings of the Royal Irish Academy. Section B: Biological, Geological, and Chemical Science, 51:53–80.

Smetacek, V. 1999. Diatoms and the ocean carbon cycle. Protist 150:25–32.

———. 2001. A watery arms race. Nature 411:745–745.

Smetacek, V., P. Assmy, and J. Henjes. 2004. The role of grazing in structuring Southern Ocean pelagic ecosystems and biogeochemical cycles. Antarctic Science 16:541–558.

Smith, L. H. and P. M. Bunje. 1999. Morphologic diversity of inarticulate brachiopods through the Phanerozoic. Paleobiology 396–408.

Sokal, R. R. and F. J. Rohlf. 1981. Biometry: the principles and practice of statistics in biological research. WH Freeman New York, 3rd ed.

Sorhannus, U. 2004. Diatom phylogenetics inferred based on direct optimization of nuclear-encoded SSU rRNA sequences. Cladistics 20:487–497.

———. 2007. A nuclear-encoded small-subunit ribosomal RNA timescale for diatom evolution. Marine Micropaleontology 65:1–12.

Spaulding, S. and M. Edlund. 2009. *Cymatopleura*. In Diatoms of the United States, http://westerndiatoms.colorado.edu/taxa/genus/Cymatopleura.

———. 2010. *Epithemia*. In Diatoms of the United States, http://westerndiatoms.colorado.edu/taxa/genus/Epithemia.

Spencer-Cervato, C. 1999. The Cenozoic deep sea microfossil record: explorations of the DSDP/ODP sample set using the Neptune database. Palaeontologia Electronica 2.

Sperling, E. A., D. Pisani, and K. J. Peterson. 2011. Molecular paleobiological insights into the origin of the brachiopoda. Evolution & Development 13:290–303.

Stanley, S. M. 1973. An explanation for Cope's rule. Evolution 1–26.

———. 1978. Chronospecies' longevities, the origin of genera, and the punctuational model of evolution. Paleobiology 26–40.

Stevens, S. S. 1946. On the theory of scales of measurement. Science 103:677–680.

Stidolph, S. R. 1985. Occurrence of the diatom *Glyphodiscus stellatus* Greville living in New Zealand coastal waters. Nova Hedwigia 41:495–504.

Stoermer, E. and T. Ladewski. 1982. Quantitative analysis of shape variation in type and modern populations of *Gomphoneis herculeana*. Nova Hedvigia, Beih 347–386.

Suto, I. 2004. Taxonomy of the diatom resting spore form genus *Liradiscus* Greville and its stratigraphic significance. Micropaleontology 50:59–79.

———. 2005. Observations on the fossil resting spore morphogenus *Peripteropsis* gen. nov. of the marine diatom genus *Chaetoceros* (Bacillariophyceae) in the Norwegian Sea. Phycologia 44:294–304.

Suto, I., R. W. Jordan, and M. Watanabe. 2009. Taxonomy of middle Eocene diatom resting spores and their allied taxa from the central Arctic Basin. Micropaleontology 55:259–312.

Suto, I., M. Watanabe, and R. W. Jordan. 2011. Taxonomy of the fossil marine diatom resting spore genus *Odontotropis*. Diatom Research 26:255–272.

Suzuki, N. and Y. Aita. 2011. Radiolaria: achievements and unresolved issues: taxonomy and cytology. Plankton and Benthos Research 6:69–91.

Swan, A. R. H. and W. B. Saunders. 1987. Function and shape in late Paleozoic (mid-Carboniferous) ammonoids. Paleobiology 297–311.

Thomas, R. D. K. and W. E. Reif. 1993. The skeleton space: a finite set of organic designs. Evolution 341–360.

Tuji, A., D. M. Williams, P. A. Sims, and Y. Tanimura. 2009. An illustrated catalogue of type specimens from the HMS Challenger voyage in Castracane's slide collection in the Natural History Museum, London. Joint Haeckel and Ehrenberg Project, Reexamination of the Haeckel and Ehrenberg Microfossil Collections as a Historical and Scientific Legacy, National Museum of Nature and Science Monographs 40:7–11.

Van Heurck, H. and W. E. Baxter. 1896. A Treatise on Diatomaceae. William Wesley & Son.

Venables, W. N. and B. D. Ripley. 2002. Modern Applied Statistics with S. Springer, New York, fourth ed. ISBN 0-387-95457-0.

Vermeij, G. 1987. Evolution and Escalation: An Ecological History of Life. Princeton University Press.

Webster, M. 2007. A Cambrian peak in morphological variation within trilobite species. Science 317:499–502.

Whittaker, R. 1960. Vegetation of the Siskiyou mountains, Oregon and California. Ecological Monographs 30:279–338.

Williams, D. M. 2007. Diatom phylogeny: Fossils, molecules and the extinction of evidence. Comptes Rendus Palevol 6:505–514.

Wills, M. 2001. Morphological disparity: a primer. *In* J. Adrain, G. Edgecombe, and B. Lieberman, eds., Fossils, phylogeny, and form: an analytical approach, 55–143. Kluwer, New York.

Wills, M., D. Briggs, and R. Fortey. 1994. Disparity as an evolutionary index: a comparison of Cambrian and Recent arthropods. Paleobiology 93–130.

Wilson, J. P. and A. H. Knoll. 2010. A physiologically explicit morphospace for tracheid-based water transport in modern and extinct seed plants. Paleobiology 36:335–355.

Witt, O. N. 1886. Über den Polierschiefer von Archangelsk-Kurojedowo im Gouv. Simbirsk. Verhandlungen der Russisch-Kaiserlichen Mineralogischen Gesellschaft zu St Petersburg 2:137–177.

Wright, S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *In* Proceedings of the Sixth International Congress on Genetics, 1:356–366.

Yanagisawa, Y. 1994. *Mediaria magna* Yanagisawa, sp. nov., a new fossil raphid diatom species useful for middle Miocene diatom biostratigraphy. Transactions and Proceedings of the Palaeontological Society of Japan. New series 174:411–425.

———. 1995a. Cenozoic diatom genus *Bogorovia* Jouse: An emended description. Transactions and Proceedings of the Palaeontological Society of Japan. New series 177:21–42.

———. 1995b. Cenozoic diatom genus *Rossiella* Desikachary et Maheshwari: An emended description. Transactions and Proceedings of the Palaeontological Society of Japan 177:1–20.

Yanagisawa, Y. and F. Akiba. 1990. Taxonomy and phylogeny of the three marine diatom genera, *Crucidenticula, Denticulopsis* and *Neodenticula*. Bulletin of the Geological Survey of Japan 41:197–301.

Zachos, J., M. Pagani, L. Sloan, E. Thomas, and K. Billups. 2001. Trends, rhythms, and aberrations in global climate 65 Ma to present. Science 292:686–693.

Zielinski, U. and R. Gersonde. 1997. Diatom distribution in southern ocean surface sediments (atlantic sector): implications for paleoenvironmental reconstructions. Palaeogeography, Palaeoclimatology, Palaeoecology 129:213–250.

A

Supplementary Figures for Chapter 3

Figure A.1 ***(following page)*****:** Metrics of morphological disparity (A-D) and taxonomic diversity (E) for the Cenozoic morphospace of marine planktonic diatoms, populated using *Neptune* database occurrences subsampled to a quota of 13 lists using by-list unweighted subsampling (UW) with 10,000 iterations. Metrics as explained in Fig. 3.2; error bars show 95% confidence intervals of subsampling.

Figure A.1: (continued)

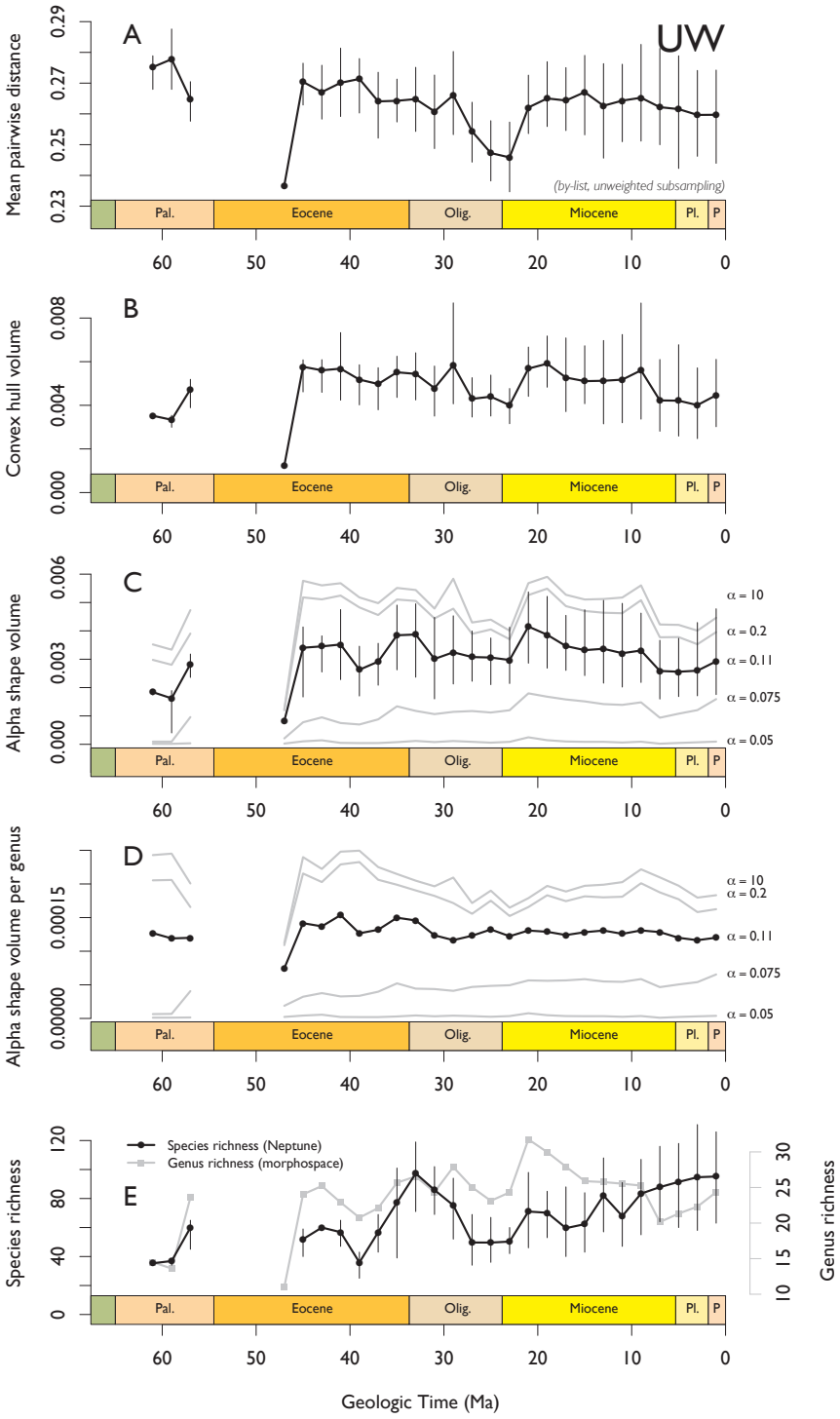
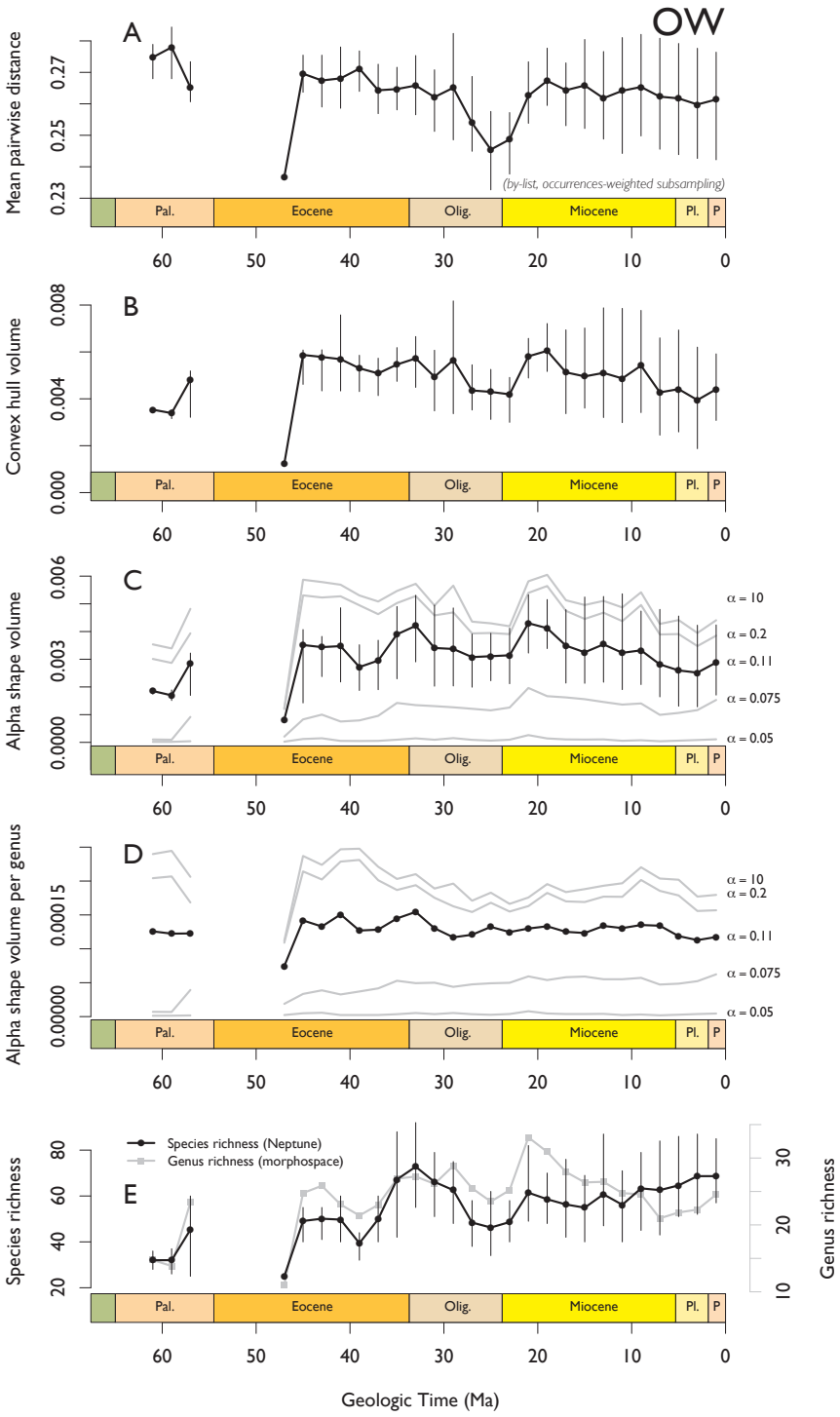


Figure A.2 *(following page)*: Metrics of morphological disparity (A-D) and taxonomic diversity (E) for the Cenozoic morphospace of marine planktonic diatoms, populated using *Neptune* database occurrences subsampled to a quota of 100 occurrences using by-list subsampling weighted by occurrences (OW) with 10,000 iterations. Metrics as explained in Fig. 3.2; error bars show 95% confidence intervals of subsampling.

Figure A.2: (continued)



B

Description of Morphospace Characters

Character number corresponds to column headings in Appendix C, character number in parentheses corresponds to index in culled dataset (as in Figures **??** and 2.2), dash indicates absence in culled dataset. Logically inapplicable characters were coded as "n", unobserved characters were coded as "?", characters observed in several states where no one state could be identified as predominant were coded as "v". Characters binary or treated as unordered multistate. Characters 1 and 2 are based on the plane shape classification published by the Systematics Association (Anonymous 1962); that reference contains a diagram explaining these shapes visually.

1 (1). Basic shape of valve outline in valve view: 0, Ellipticus; 1, Oblongus vel rectangularis; 2, Rhombicus; 3, Ovatus; 4, Triangularis.

2 (2). Aspect ratio of shape of valve outline in valve view: 0, Linearis; 1, Anguste; 2, Anguste late; 3, Late; 4, 1:1; 5, Latissime; 6, Depresse; 7, Perdepresse.

3 (3). Additional undulations or crenulations superimposed on basic outline in valve view: 0, None; 1, One central expansion; 2, One eccentric expansion; 3, One central constriction; 4, Two constrictions; 5, Three constrictions; 6, Four or more constrictions.

4 (4). Torsion of the frustule along the longest axis in the valvar plane (in pennate diatoms, the apical axis): 0, No torsion (valvar plane flat); 1, Torsion (valvar plane twisted).

5 (5). Curvature of the frustule along the longest valvar-plane axis in the valvar plane, in valve view (also termed dorsoventrality): 0, No curvature; 1, Constant curvature along axis; 2, Axis mostly straight, curved at center only, axis slightly bent; 3, Axis mostly straight, curved

at center only, axis strongly bent; 4, Axis bent in opposite directions either side of the midline (i.e. cell sigmoidal in valve view).

6 (-). Torsion of the frustule along the pervalvar axis: 0, No torsion; 1, Torsion.

7 (6). Curvature of the frustule along the pervalvar axis: 0, No curvature (pervalvar axis straight); 1, Curvature (pervalvar axis not straight).

8 (7). Curvature of the frustule along the longest valvar-plane axis in the apical plane, in girdle view (in pennate diatoms, curvature of the apical axis in the apical plane): 0, No curvature; 1, Constant curvature along axis; 2, Axis mostly straight, curved at center only, axis slightly bent; 3, Axis mostly straight, curved at center only, axis strongly bent; 4, Sinuous or sigmoidal.

9 (8). Plan-view shape of apices (if heteropolar, shape of head-pole): 0, Rostrate; 1, Capitate; 2, Apiculate; 3, Broadly rounded; 4, Acute or acutely rounded; 5, Long needle-like horn longer than main body

10 (-). Plan-view shape of apex at foot-pole, if heteropolar: 0, Rostrate; 1, Capitate; 2, Apiculate; 3, Broadly rounded; 4, Acute or acutely rounded; 5, Long needle-like horn longer than main body.

11 (9). Valve similarity: 0, Valves similar (isovalvar); 1, Valves dissimilar (heterovalvar).

12 (10). General topography of valve face (ignoring apical elevations, if any): 0, Flat; 1, Convex, diameter of curvature significantly greater than the shortest dimension of the valve in the valvar plane; 2, Convex, diameter of curvature comparable to shortest dimension of valve in valvar plane; 3, Concave; 4, Valve face set in deeply invaginated interior; 5, Strongly drawn out vertically into a cone-like shape.

13 (11). Second order topography of folds, superimposed on that above: 0, None; 1, Concentric undulations/rises; 2, Transverse plications/folds or undulations, gentle; 3, Transverse plications/folds or undulations, strong; 4, Corrugations; 5, Longitudinal plications (i.e. parallel to apical axis).

14 (-). Topographic sectors (alternately raised and depressed): 0, None; 1, Three, five, or up to ten sector pairs.

15 (-). Sulci dividing valve face into segments (per elongation axis, e.g. 3 sulci in triangular form = 1 sulcus): 0, Absent; 1, One sulcus (two segments); 2, Two sulci (three segments); 3, Three sulci; 4, Four sulci; 5, Five sulci.

16 (-). Orientation of sulci: 0, Sulci transverse to axis of elongation, i.e. apical axis in pennates, longest axis in centrics; 1, Sulci longitudinal, i.e. parallel to axis of elongation, i.e. apical axis in pennates, longest axis in centrics.

17 (-). Depth of sulci: 0, Shallow (less than mantle height); 1, Deep.

18 (12). Topography on apices (or, for centric forms without clear apices, along valve face margin): 0, No topography (flat); 1, Slightly elevated, height less than mantle height; 2, Pronounced apical elevation, height greater than mantle height but less than longest valve width in valve view; 3, Extreme apical elevation, height greater than longest valve width in valve view ("apical horns".

19 (-). Similarity in topography between or among apices: 0, Topography similar between or among apices; 1, One apex more elevated than other(s).

20 (13). Width of apical elevations: 0, Less than the full width of the apices; 1, Apical elevations as wide as the apices.

21 (14). Shape of apical elevation summit: 0, Pointed; 1, Flattened; 2, Rounded.

22 (15). Central elevation or depression: 0, Absent; 1, Elevation present, comparable in height to smallest dimension of valve in valvar plane; 1, Elevation present, significantly larger in height than smallest dimension of valve in valvar plane; 2, Depression present.
23 (16). Central elevation or depression position: 0, Located at center of valve; 1, Located in eccentric position.
24 (17). Central elevation shape: 0, Narrow, concave ("horn" type); 1, Wide, convex ("dome" type).
25 (18). Central elevation curvature: 0, Central elevation straight; 1, Central elevation curved.
26 (19). Angle between valve face and mantle: 0, No clear distinction; 1, 90 degree angle; 2, Obtuse angle.
27 (20). Ornament at junction between valve face and mantle: 0, No ornament; 1, Row of simple, short spines or spinules; 2, Numerous elongated or ornamented spines or conspicuous marginal processes, can be setae or exits of rimoportulae.
28 (21). Evidence for ornament at junction between valve face and mantle, or marginal ridge, providing linking structure: 0, Absent; 1, Interlocking spines (or setae).
29 (22). Marginal ridge at junction between valve face and mantle: 0, Absent; 1, Thickened rim; 2, Low raised vertical ridge not rising above apical elevations where present; 3, High raised vertical ridge, extending to height of apical elevations where present.
30 (23). Variation in height of marginal ridge around perimeter: 0, Height constant; 1, Marginal ridge lower in places, at apices if present.
31 (24). Depth of mantle: 0, No mantle apparent; 1, Shallow, mantle depth less than half of the shortest dimension of the valve face in valve view (for circular forms: mantle depth less than radius); 2, Deep, mantle depth roughly equal to or greater than half the shortest dimension (radius for circular forms), but not greater than the shortest dimension (diameter for circular forms); 3, Very deep, mantle depth greater than the shortest dimension of the valve face in valve view (for circular forms: mantle depth greater than diameter).
32 (25). Symmetry of mantle in girdle view: 0, Mantle depth equal around valve; 1, Mantle deeper on one side than on the other.
33 (26). Mantle thickness (relative to valve face): 0, Thin; 1, Normal; 2, Thick.
34 (27). Mantle shape in cross section: 0, Straight; 1, Convex ('bowed out'); 2, Concave
35 (28). Warts or plaques on mantle exterior: 0, Absent; 1, Present.
36 (29). Circumferential ridge or ridges on mantle exterior: 0, Absent; 1, Present.
37 (30). Spines on mantle exterior: 0, Absent; 1, Present.
38 (31). Ornament at mantle edge (furthest from valve face): 0, None, mantle edge plain; 1, Rim (annular thickening surrounding mangle edge); 2, Ribbed (radial ridges on mantle edge); 3, Crimped; 4, Vertical marginal spines; 5, Long, horizontally projecting spines.
39 (32). Brim (mantle edge turned outward like a hat): 0, Absent; 1, Narrow brim present; 2, Wide, upturned/scalloped brim.
40 (-). Brim ornament: 0, None; 1, Spines.
41 (33). Mantle pore type: 0, No pores; 1, Pores of unknown structure (puncta); 2, Poroid areaolae; 3, Loculate or pseudoloculate areolae; 4, Simple perforations (unoccluded on either side).
42 (34). Mantle pore arrangement: 0, In (vertical) columns; 1, In horizontal rows; 2, In diagonal rows; 3, In small fields or patches; 4, Irregularly scattered.
43 (35). External costae/ribs/ridges: 0, Absent; 1, Costae in reticulate pattern; 2, Anastomosing costae, covering external valve surface 3, Costae separating rows of pores; 4,

Prominent radial costae in central area only; 5,Short costae at valve edge only.
44 (36). Rays: 0, Absent; 1, Present.
45 (-). Shape of rays: 0, Spade-shaped; 1, Drop- or pear-shaped; 2, Fan-shaped; 3, Ovate or sub-ovate.
46 (37). One ray differs from the others: 0, No, rays similar; 1, Yes.
47 (-). Rays raised: 0, No, rays not raised; 1, Rays domed or arched.
48 (38). Ray slits (underside of ray chamber): 0, Slits pierce floor of chamber over central expansion only; 1, Slits pierce floor of chamber over central expansion and at least a part of the marginal prolongation.
49 (39). Granules or warts: 0, Absent; 1, Present on valve face; 2, In radiating pattern; 3, In central area only.
50 (40). Surface texture: 0, None; 1, Corrugated; 2, Pitting in between areolae; 3, Radial ridges and grooves.
51 (41). Distinct central area (other than a sternum): 0, None; 1, Distinct central area, round shape; 2, Distinct central area, naviculoid shape.
52 (42). Distinguishing characteristic of central area: 0, Finer or coarser pores or striae; 1, Hyaline; 2, Distant and scattered pores.
53 (43). Ornament at periphery of central area: 0, None; 1, Hyaline, non-areolated sulcus; 2, Ring of spines or tubercles; 3, Decorated ring of triangular pustulae; 4, Corona (ring of irregular spines).
54 (-). Setae: 0, None; 1, One near each pole; 2, Many arising around periphery.
55 (44). Tubular spines: 0, Absent; 1, One or two tubular spines at center of valve; 2, Ring of tubular spines.
56 (45). Spinules: 0, Absent; 1, Present between pores.
57 (-). Hair-like filaments: 0, Absent; 1, Present.
58 (46). Collar (extended membranous wing on the outer side of the valve) or Carina (flat, collar-like structure between central area and valve face margin): 0, Absent; 1, Present.
59 (47). Robust linking processes that slot together, clasp each other, or interdigitate: 0, Absent; 1, Present on apices/on or as apical elevations; 2, Present at valve center or in central area.
60 (48). Shape of structural pattern center of primary silica ribs: 0, Ring-shaped principal rib (annulus); 1, Linear principal rib (sternum).
61 (49). Packing/coordination of pores: 0, Hexagonal; 1, Square; 2, In rows; 3, Scattered irregularly.
62 (50). If hexagonal, arrangement of pores: 0, In straight rows (for centrics, lineata-type tangential areolation; for pennates, decussate-punctate and tranverse-oblique striate); 1, In straight rows, but collected in radial bundles (radial fasciculate); 2, In curved rows, collected in radial bundles with curved edges (radial fasciculate, curvatulus type); 3, With secondary rows in spirals; 4, In rows concave towards margin (eccentrica type).
63 (51). If square or in rows, orientation of pore rows relative to structural pattern center: 0, Orthogonal to structural pattern center/sternum (for centrics, radial areolation, for pennates, transverse and longitudinal striae); 1, Orientation variable along pattern center; 2, Orientation variable along pattern center, but perpendicular to apical axis/axis of elongation.
64 (52). If arrangement of pores variable along pattern center, angle with pattern center in the middle of the diatom: 0, Orthogonal; 1, Divergent (radiating); 2, Convergent.

65 (-). If arrangement of pores variable along pattern center, angle with pattern center at the apices of the valve: 0, Orthogonal; 1, Divergent (radiating); 2, Convergent.
66 (53). If pores in rows ("striae" in pennates), number of rows per striation: 0, One (in pennates, uniseriate), 1, Two (in pennates, biseriate); 2, Three or more (in pennates, multiseriate).
67 (54). Uniformity of pore size: 0, Uniform/homogeneous; 1, Larger at pattern center, becoming smaller toward margin; 2, Smaller at pattern center, becoming larger toward margin; 3, Irregular
68 (55). Pore size: 0, Fine; 1, Normal; 2, Coarse; 3, Very large (with a diameter a quarter of the shortest radius of the valve face, or larger).
69 (56). Pore shape: 0, Circular; 1, Oval; 2, Quadrate; 3, Rectangular (but not quadrate); 4, Quadrangular (but not rectangular); 5, Slit-like; 6, Papillose.
70 (57). Average proximity of pores: 0, Closely packed (edges of pores less than one pore radius apart); 1, Loosely packed (edges of pores more than one pore radius but not much more than one pore diameter apart); 2, Isolated pores (edges of pores more than one pore diameter apart).
71 (58). Regularity of pore proximity/spacing: 0, Regular; 1, Irregular, more closely spaced at poles; 2, Irregular, more widely spaced at poles; 3, Irregular, more closely spaced at margins; 4, Irregular, more widely spaced at margins.
72 (59). Pore openings at external valve surface: 0, No constriction; 1, Foramen; 2, Velum.
73 (-). Pore openings at internal valve surface: 0, No constriction; 1, Foramen; 2, Velum.
74 (-). Type of velum: 0, Cribrum or hymen; 1, Rota; 2, Vola; 3, Bars.
75 (-). Velum topography: 0, Velum flush with valve surface; 1, Depressed velum; 2, Domed or raised velum.
76 (60). Pore structure identified as pseudoloculate: 0, No; 1, Yes.
77 (61). Alveoli: 0, Absent; 1, Present.
78 (62). Porelli in between pores: 0, Absent; 1, Present.
79 (63). Passage pores: 0, Absent; 1, Present.
80 (64). Bullulae: 0, Absent; 1, Present.
81 (65). Pseudonodulus: 0, Absent; 1, Present.
82 (66). Ring of specialized openings: 0, Absent; 1, Ring of radially elongate openings; 2, Ring of subtriangular apertures.
83 (67). Hyaline area at margin of valve face: 0, Absent; 1, Present.
84 (68). Hyaline ring near margin of valve face: 0, Absent; 1, Present.
85 (69). Apical pseudoseptum (membranous costa on the inner side of the valve projecting in the valvar plane): 0, Absent; 1, Present.
86 (70). Annular pseudoseptum (diaphragm-like ingrowth of cell wall projecting into cell interior) on mantle. "Ringleiste": 0, Absent; 1, Present at or very near mantle edge (farthest from valve face); 2, Present at or very near mantle edge (farthest from valve face), but at apices only; 3, Present between mantle edge and rim.
87 (71). Pseudoseptae parallel to the transapical plane or along a radius (for non-bilateral forms): primary pseudoseptae (sensu Yanagisawa and Akiba, 1990; internal ingrowth of the valve that spans the valve from side to side, penetrates deeply separating valve completely into compartments, also referred to as costae): 0, Absent; 1, Present, without marginal thickenings or branching; 2, Present, with marginal thickenings; 3, Present and branching.
88 (72). Crossbars or basal ridges atop primary pseudoseptae: 0, Absent; 1, Present.

89 (73). Pseudoseptae parallel to the transapical plane or along a radius: secondary pseudoseptae (sensu Yanagisawa and Akiba, 1990; internal ingrowth of the valve that spans the valve from side to side, but does not penetrate deeply and does not separate valve completely into compartments; also referred to as costae or ribs): 0, Absent; 1, Present but not associated with striae, or dividing striae into groups or sectors; 2, Present between each stria (=transverse ribs in pennates).
90 (74). Marginal ribs (ingrowth projecting into valve interior parallel to the transapical plane or along a radius, but not spanning the valve side to side): 0, Absent; 1, Present.
91 (75). Apical fields (if heteropolar, on head-pole): 0, Absent/undifferentiated from rest of valve face; 1, Ocelli (plate of silica, normally with thickened, structureless rim, pierced by closely packed holes or porelli); 2, Pseudocelli (field of pores of smaller size than the main part of the valve face); 3, Ocelluli (structured the same as an ocellus, but with few porelli and a raised rim); 4, Ocellulimbus (on the mantle, composed of rows of small pores arranged in an ordered manner and depressed from the valve surface, Siver et al 2006).
92 (76). Apical pore field pore arrangement: 0, Radial (or scattered); 1, In rows.
93 (77). Apical field costae (if heteropolar, on head-pole): 0, Absent; 1, Costate apical field, apical field composed of lamellae or bars with pores sunk between each bar.
94 (78). Apical fields on foot-pole, if heteropolar: 0, Absent/undifferentiated from rest of valve face; 1, Ocelli; 2, Pseudocelli; 3, Ocelluli; 4, Apical area hyaline.
95 (-). Apical field costae (if heteropolar, on foot-pole): 0, Absent; 1, Costate apical field, apical field composed of lamellae or bars with pores sunk between each bar.
96 (79).Apical ornament: 0, Absent; 1, Single linking spine at each apex (projecting parallel to pervalvar axis); 2, Two linking spines at each apex (projecting parallel to pervalvar axis); 3, Two prominent spines on one pole (projecting perpendicular to pervalvar axis).
97 (-). Pores on sides of apical elevation: 0, Continuous with/similar to rest of valve face surface; 1, No pores on sides of apical elevation; 2, Pores present, but significantly reduced in number relative to valve surface.
98 (80). Strutted processes (fultoportulae): 0, Absent; 1, One present; 2, Two present; 3, Three or more present.
99 (81). Location of strutted process(es): 0, At center of valve face; 1, On valve face, not in center; 2, At rim (valve face/margin junction); 3, In ring around mantle; 4, In ring beneath processes.
100 (82). Structure of strutted process(es): 0, Short tubes; 1, Long tubes.
101 (83). Labiate processes (rimoportulae): 0, Absent; 1, One present; 2, Two Present; 3, Three or more present.
102 (84). Location of labiate process(es): 0, At center of valve face; 1, On valve face, not in center; 2, At rim (valve face/margin junction); 3, On mantle; 4, Along sternum; 5, Randomly scattered; 6, In apices.
103 (85). External opening of labiate process(es): 0, Simple; 1, With external tube; 2, Slightly raised; 3, Projecting with cap, slits or barbs.
104 (-). Internal opening of labiate process(es): 0, On stalk; 1, Not stalked; 2, Stalked, straight; 3, Stalked, curved.
105 (-). Shape of internal opening of labiate process(es): 0, Simple; 1, Crimped; 2, Kidney-shaped; 3, Complex.
106 (-). Macrorimoportulae: 0, Absent; 1, Present.
107 (-). Sternum number: 0, One; 1, Two.

108 (86). Hyaline axial area width: 0, Narrow; 1, Wide, expanded; 2, No discernible/distinct axial hyaline area.
109 (87). Sternum location: 0, Central/axial; 1, Eccentric/lateral on valve face; 2, On rim; 3, On mantle.
110 (88). Sternum shape: 0, Constant width along length; 1, Widens at poles; 2, Narrows at poles.
111 (89). Fascia: 0, Absent; 1, Present.
112 (90). Raphe location: 0, Absent; 1, Central/axial; 2, Eccentric/lateral on valve face; 3, On rim; 4, On mantle.
113 (91). Number of branches of raphe slit: 0, One; 1, Two.
114 (92). Raphe extent: 0, From pole to pole; 1, Around entire valve circumference.
115 (93). Raphe sinuosity: 0, Raphe linear; 1, Raphe sigmoidal; 2, Raphe sinuous or curved.
116 (94). Conopeum: 0, Absent; 1, Present.
117 (95). External polar raphe endings: 0, No terminal fissure, plain ending; 1, Transverse terminal fissure (T-shaped); 2, Double or forked; 3, Straight terminal fissure; 4, Slightly deflected terminal fissure; 5, Bent terminal fissure; 6, Hooked terminal fissure.
118 (96). External central raphe endings: 0, Straight, simple; 1, Straight, expanded (e.g. to form a central pore); 2, T-shaped; 3, Forked; 4, Deflected, bent or hooked in the same direction; 5, Deflected, bent or hooked in opposite directions.
119 (-). Central nodule: 0, Not transapically expanded; 1, Transapically expanded (i.e. a stauros).
120 (97). Raphe canal: 0, Absent; 1, Present.
121 (98). Raphe keel: 0, Absent; 1, Present.
122 (99). Fibulae: 0, Absent; 1, Present.
123 (100). Relative thickness of raphe sides: 1, Both sides of raphe evenly thick; 1, One side of raphe thicker than other (raphe opens laterally).

C

Morphospace Data Matrix

Table C.1: Morphospace data matrix (characters 1-25)

Genus↓ *Character* →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Abas	1	4	0	0	0	0	0	0	n	n	0	1	0	0	0	n	n	3	0	0	n	0	n	n	n
Achnanthes	1	1	3	0	0	0	0	2	4	n	1	1	0	0	0	n	n	0	n	n	n	0	0	0	0
Actinocyclus	0	4	0	0	0	0	0	0	n	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Actinoptychus	0	4	0	0	0	0	0	0	n	n	0	0	0	1	0	n	n	0	n	n	n	0	n	n	n
Amblypyrgus	0	4	0	0	0	0	0	0	n	n	0	2	0	0	0	n	n	0	n	n	n	0	n	n	n
Amphora	0	1	0	0	1	0	0	1	v	n	0	2	0	0	0	n	n	0	0	n	n	0	n	n	n
Anaulus	0	1	0	0	0	0	0	0	3	n	0	2	0	0	0	n	n	1	0	0	0	0	n	n	n
Ancylopyrgus	0	4	0	0	0	0	1	0	n	n	0	1	0	0	0	n	n	0	n	n	n	1	1	1	n
Annellus	0	4	0	0	0	0	0	0	n	n	0	4	0	0	0	n	n	0	n	n	n	0	n	n	n
Arachnoidiscus	0	4	0	0	0	0	0	0	n	n	1	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Archepyrgus	0	4	0	0	0	0	0	0	n	n	0	1	0	0	0	n	n	1	n	n	n	0	n	n	n
Asterolampra	0	4	0	0	0	0	0	0	n	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Asteromphalus	0	4	0	0	0	0	0	0	n	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Aulacodiscus	0	4	0	0	0	0	0	0	n	n	0	0	0	0	0	n	n	1	n	n	n	0	n	n	n
Aulacoseira	0	4	0	0	0	0	0	0	n	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Azpeitia	0	4	0	0	0	0	0	0	n	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Bacillaria	0	0	0	0	0	0	0	0	0	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Bacteriastrum	0	4	0	0	0	0	0	0	n	n	0	0	0	0	0	n	n	0	n	n	n	3	0	n	n
Bacterosira	0	4	0	0	0	0	0	0	n	n	0	0	0	0	0	n	n	n	n	n	n	3	0	n	n
Basilicostephanus	0	4	0	0	0	0	1	0	n	n	0	1	0	0	0	n	n	0	n	n	n	v	n	n	n
Baxteriopsis	0	1	1	0	0	0	0	0	1	n	0	0	0	0	0	n	n	2	0	1	1	1	0	1	0
Biddulphia	0	2	6	0	0	0	0	0	0	n	0	2	2	0	4	0	1	2	0	1	2	1	0	1	0
Bilingua	0	1	1	0	0	0	0	0	3	n	0	0	0	0	0	n	n	1	0	1	1	1	0	1	0
Bogorovia	0	1	0	0	0	0	0	0	0	n	1	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Brightwellia	0	4	0	0	0	0	0	0	n	n	0	1	0	0	n	n	n	0	n	n	n	0	n	n	n
Caloneis	1	1	0	0	0	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Campyloneis	0	3	0	0	0	0	0	0	n	n	1	3	0	0	0	n	n	0	n	n	n	0	n	n	n

Table C.1: (continued)

Genus↓ *Character* →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Ceratoneis	1	1	1	1	1	0	0	0	5	n	0	2	0	0	0	n	n	0	0	n	n	0	n	n	n
Cereataulus	0	4	0	1	0	0	0	0	n	n	0	1	0	0	0	n	n	1	n	n	n	0	n	n	n
Cestodiscus	0	4	0	0	0	0	0	0	n	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Chaetoceros	0	3	0	0	0	0	0	0	3	n	0	1	0	0	0	n	n	1	0	0	1	1	0	1	0
Charcotia	0	4	0	0	0	0	0	0	n	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Cladogramma	0	4	0	0	0	0	0	0	n	n	?	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Clavicula	1	0	0	0	0	0	0	0	1	n	0	?	?	0	0	n	n	?	?	?	?	?	?	?	?
Cocconeis	0	3	0	0	0	0	0	0	3	n	1	3	0	0	n	n	n	0	n	n	n	0	n	n	n
Corethron	0	4	0	0	0	0	0	0	n	n	1	2	0	0	0	n	n	0	n	n	n	0	n	n	n
Coscinodiscus	0	4	0	0	0	0	0	0	n	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Cosmiodiscus	0	4	0	0	0	0	0	0	n	n	0	2	0	0	0	n	n	0	n	n	n	0	n	n	n
Craspedodiscus	0	4	0	0	0	0	0	0	n	n	0	0	1	0	0	n	n	0	n	n	n	3	0	n	n
Crucidenticula	1	2	0	0	0	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Cussia	0	1	0	0	0	0	0	0	4	n	?	?	?	0	0	n	n	?	?	?	?	?	?	?	?
Cyclotella	0	4	0	0	0	0	0	0	n	n	0	1	1	0	0	n	n	0	n	n	n	3	1	n	n
Cymatodiscus	0	3	0	0	0	0	0	2	3	n	1	?	1	0	0	n	n	0	n	n	n	3	0	n	n
Cymatogonia	4	4	0	0	0	0	0	0	3	n	0	0	0	1	0	n	n	0	n	n	n	0	n	n	n
Cymatopleura	1	1	3	0	0	0	0	0	4	n	0	0	2	0	0	n	n	0	n	n	n	0	n	n	n
Cymatosira	0	1	1	0	0	0	0	0	3	n	1	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Cymatotheca	0	3	0	0	0	0	0	2	3	n	1	?	2	0	0	n	n	0	n	n	n	0	0	n	n
Cymbella	0	1	0	0	1	0	0	0	v	n	0	0	0	0	0	n	n	0	n	n	0	n	n	n	n
Dactyliosolen	0	4	0	0	0	0	0	0	n	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Delphineis	0	v	0	0	0	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Denticula	0	2	0	0	0	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Denticulopsis	1	1	0	0	0	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Dextradonator	0	3	0	0	0	0	0	0	3	n	1	0	0	0	0	n	n	3	0	0	0	0	n	n	n
Diatoma	0	2	0	0	0	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n

Table C.1: (continued)

Genus↓ *Character* →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Dimerogramma	0	1	0	0	0	0	0	0	3	n	0	0	2	0	0	n	n	0	0	n	n	0	n	n	n
Diploneis	0	2	3	0	0	0	0	0	3	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Discodiscus	0	4	0	0	0	0	0	0	n	n	0	1	0	0	0	n	n	0	n	n	n	3	0	n	n
Endictya	0	4	0	0	0	0	0	0	n	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Epithemia	3	7	0	0	1	0	0	0	0	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Ethmodiscus	0	4	0	0	0	0	0	0	n	n	?	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Eucampia	0	1	0	0	0	0	0	0	3	n	0	1	0	0	0	n	n	2	1	0	1	0	n	n	n
Eunotia	1	1	v	0	1	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Eunotogramma	0	1	0	0	1	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Fenestrella	0	4	0	0	0	0	0	0	n	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Fragilaria	0	1	0	0	0	0	0	0	0	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Fragilariopsis	1	1	0	0	0	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Gladiopsis	0	4	0	0	0	0	0	0	n	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Gladius	0	4	0	0	0	0	0	0	n	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Glyphodiscus	0	4	0	0	0	0	0	0	n	n	1	1	1	0	0	n	n	1	0	0	1	1	0	1	0
Gomphonema	0	1	1	0	0	0	0	0	3	3	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Grammatophora	1	2	0	0	0	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Grunowiella	0	0	0	0	0	0	0	0	1	3	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Gyrosigma	0	1	0	0	4	0	0	0	4	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Hemiaulus	0	2	0	0	0	0	0	0	3	n	0	2	0	0	0	n	n	3	0	0	0	0	n	n	n
Hemidiscus	3	6	0	0	0	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Horodiscus	0	4	0	0	0	0	0	0	n	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Huttonia	0	2	0	0	0	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Hyalodiscus	0	4	0	0	0	0	0	0	n	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Ikebea	0	1	0	0	0	0	0	0	3	0	0	1	2	0	0	n	n	0	n	n	n	0	n	n	n
Katathiraia	0	2	0	0	0	0	0	0	3	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Kerkis	0	2	0	0	0	0	0	0	3	n	0	0	3	0	0	n	n	1	0	1	0	0	n	n	n

Table C.1: (continued)

Genus↓ *Character* →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Kisseleviella	0	2	1	0	0	0	0	0	4	n	0	0	0	0	0	n	n	0	n	n	n	1	0	1	0
Kozloviella	3	6	0	0	0	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	1	0	1	0
Kreagra	0	4	0	0	0	0	0	0	n	n	0	2	0	0	0	n	n	0	n	n	n	0	n	n	n
Liriogramma	0	2	1	0	0	0	0	0	3	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Lisitzinia	1	4	3	0	0	0	0	0	0	n	0	1	0	0	4	0	1	0	n	n	n	0	n	n	n
Lithodesmium	4	4	4	0	0	0	0	0	2	n	0	1	1	0	0	0	0	1	0	1	0	0	n	n	n
Mastogloia	0	2	0	0	0	0	0	0	2	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Mediaria	0	1	0	0	0	0	0	0	4	n	1	1	6	0	0	n	n	0	n	n	n	0	n	n	n
Melosira	0	4	0	0	0	0	0	0	n	n	0	v	0	0	0	n	n	n	n	n	n	0	n	n	n
Microorbis	0	4	0	0	0	0	0	0	n	n	1	1	0	0	0	n	n	0	n	n	n	v	v	n	n
Monobrachia	0	?	0	0	0	0	v	0	?	?	0	5	0	0	0	n	n	0	n	n	n	2	0	0	0
Navicula	0	1	0	0	0	0	0	0	0	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Neobrunia	0	4	0	0	0	0	0	0	n	n	0	0	1	0	0	n	n	0	n	n	n	0	n	n	n
Neodelphineis	0	0	0	0	0	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Neodenticula	1	1	0	0	0	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Nitzschia	0	1	0	0	0	0	0	0	0	n	0	0	5	0	0	n	n	0	n	n	n	0	n	n	n
Odontella	0	2	0	0	0	0	0	0	3	n	0	2	0	0	0	n	n	1	0	v	v	v	0	1	0
Opephora	3	1	0	0	0	0	0	0	3	3	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Paralia	0	4	0	0	0	0	0	0	n	n	1	0	0	0	0	n	n	0	n	n	n	0	0	n	n
Peponia	0	3	1	0	0	0	0	0	2	n	0	1	1	0	0	n	n	1	0	0	2	0	n	n	n
Plagiogramma	0	1	0	0	0	0	0	0	3	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Planktoniella	0	4	0	0	0	0	0	0	n	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Pleurosigma	2	1	0	0	4	0	0	0	3	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Podosira	0	4	0	0	0	0	0	0	n	n	0	2	0	0	0	n	n	0	n	n	n	0	n	n	n
Porosira	0	4	0	0	0	0	0	0	n	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Praethalassiosiropsis	0	4	0	0	0	0	0	0	n	n	0	0	1	0	0	n	n	0	n	n	n	0	n	n	n
Pseudodimerogramma	1	1	0	0	0	0	0	0	3	3	0	?	?	0	0	n	n	?	?	?	?	?	?	?	?

Table C.1: (continued)

Genus↓ *Character* →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Pseudoeunotia	0	1	0	0	2	0	0	0	3	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Pseudopodosira	0	4	0	0	0	0	0	0	n	n	2	2	0	0	0	n	n	0	n	n	n	v	0	1	0
Pseudorutilaria	1	0	1	0	0	0	0	0	3	n	0	0	2	0	0	n	n	2	0	1	1	1	0	1	0
Pseudostictodiscus	0	3	0	0	0	0	0	0	n	n	0	0	1	0	0	n	n	0	n	n	n	0	n	n	n
Pseudotriceratium	4	4	0	0	0	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Pyrgupyxis	0	4	0	0	0	0	0	0	n	n	0	2	0	0	0	n	n	0	n	n	n	1	0	0	1
Pyxilla	0	4	0	0	0	0	0	0	n	n	1	2	0	0	0	n	n	0	n	n	n	2	0	0	0
Rattrayella	0	4	0	0	0	0	0	0	n	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Rhabdonema	0	1	0	0	0	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Rhaphidodiscus	0	3	0	0	0	0	0	0	3	n	0	1	0	0	0	n	n	0	n	n	n	3	0	n	n
Rhaphoneis	0	2	0	0	0	0	0	0	2	n	0	0	0	0	n	n	n	0	n	n	n	0	n	n	n
Rhizosolenia	0	4	0	0	0	0	0	0	n	n	0	5	0	0	0	n	n	0	n	n	n	2	1	0	0
Rhoicosphenia	3	1	0	0	0	0	0	1	3	3	1	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Rhopalodia	3	6	0	0	1	0	0	0	v	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Rhynchopyxis	0	4	0	0	0	0	0	0	n	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Riedelia	0	2	0	0	0	0	0	0	3	n	0	2	0	0	0	n	n	3	0	0	0	0	n	n	n
Rocella	0	4	0	0	0	0	0	0	n	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Roperia	0	4	0	0	0	0	0	0	n	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Rossiella	0	1	0	0	0	0	0	0	0	n	1	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Rouxia	0	1	0	0	0	0	0	4	2	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Rutilaria	0	2	1	0	0	0	0	0	0	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Rylandsia	0	4	0	0	0	0	0	0	n	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Sceptroneis	1	0	1	0	0	0	0	0	1	3	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Simonseniella	0	3	0	0	0	1	1	0	n	n	0	5	0	0	0	n	n	n	n	n	n	2	0	0	1
Skeletonema	0	4	0	0	0	0	0	0	n	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Sphynctolethus	0	2	0	0	0	0	0	0	3	n	0	1	0	0	2	0	0	1	0	0	1	0	n	n	n
Stellarima	0	4	0	0	0	0	0	0	n	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n

Table C.1: (continued)

Genus↓ *Character* →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Stephanodiscus	0	4	0	0	0	0	0	0	n	n	1	0	1	0	0	n	n	0	n	n	n	0	n	n	n
Stephanogonia	0	4	0	0	0	0	0	0	n	n	0	2	0	0	0	n	n	0	n	n	n	1	0	1	0
Stephanopyxis	0	4	0	0	0	0	0	0	n	n	0	2	0	0	0	n	n	0	n	n	n	0	n	n	n
Stictodiscus	4	4	1	0	0	0	0	0	2	n	0	0	0	0	0	n	n	0	n	n	0	n	n	n	n
Strangulonema	0	4	0	0	0	0	0	0	n	n	0	1	0	0	0	n	n	0	n	n	n	0	n	n	n
Surirella	0	2	0	0	0	0	0	0	4	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Synedra	1	0	0	0	0	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Tetracyclus	0	2	1	0	0	0	0	0	1	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Thalassionema	1	0	0	0	0	0	0	0	3	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Thalassiosira	0	4	0	0	0	0	0	0	n	n	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Thalassiothrix	1	0	0	0	0	0	0	0	3	4	0	0	0	0	0	n	n	0	n	n	n	0	n	n	n
Trachyneis	0	1	0	0	0	0	0	0	4	n	0	2	0	0	0	n	n	0	n	n	n	0	n	n	n
Triceratium	4	4	0	0	0	0	0	0	4	n	0	1	0	0	0	n	n	1	0	1	2	0	n	n	n
Trinacria	4	4	0	0	0	0	0	0	4	n	0	1	0	0	0	n	n	2	0	1	1	1	0	1	0
Trochosira	0	4	0	0	0	0	0	0	n	n	0	1	0	0	0	n	n	0	n	n	n	0	0	n	n
Trochus	0	4	0	0	0	0	0	0	n	n	1	0	0	0	0	n	n	0	n	n	n	1	0	1	0
Tropidoneis	0	1	0	0	0	0	0	0	4	n	0	1	5	0	0	n	n	0	n	n	n	0	n	n	n

Table C.2: Morphospace data matrix (characters 26-48)

Genus↓ *Character* →	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Abas	1	1	1	1	0	1	0	1	0	0	0	0	0	0	n	1	4	0	0	n	n	n	n
Achnanthes	1	0	n	0	n	2	0	1	1	0	0	0	0	0	n	2	0	0	0	n	n	n	n
Actinocyclus	1	0	n	1	0	1	0	1	0	1	0	0	0	0	n	1	?	0	0	n	n	n	n
Actinoptychus	2	1	0	1	0	1	0	1	0	1	0	1	1	0	n	?	0	0	0	n	n	n	n
Amblypyrgus	0	0	n	0	n	3	0	1	0	0	0	0	2	0	n	3	2	0	0	n	n	n	n
Amphora	v	0	n	n	n	1	1	1	1	0	0	0	0	0	n	v	v	v	0	n	n	n	n

Table C.2: (continued)

Genus↓ *Character* →	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Anaulus	0	0	n	0	n	3	0	1	1	0	0	0	0	0	n	2	0	0	0	n	n	n	n
Ancylopyrgus	2	0	n	2	0	3	0	1	0	0	0	0	0	0	n	3	2	0	0	n	n	n	n
Annellus	1	0	n	0	n	2	0	1	0	0	0	0	0	0	n	3	0	1	0	n	n	n	n
Arachnoidiscus	1	0	n	1	0	1	0	1	0	0	0	0	1	0	n	2	0	0	0	n	n	n	n
Archepyrgus	1	2	1	2	1	3	0	1	0	0	0	0	0	0	n	2	0	4	0	n	n	n	n
Asterolampra	0	0	n	0	n	0	0	1	1	0	0	0	0	0	n	3	2	0	1	0	0	1	1
Asteromphalus	0	0	n	0	n	0	0	1	1	0	0	0	0	0	n	3	0	v	1	1	1	1	1
Aulacodiscus	2	2	0	0	n	1	0	1	0	0	0	0	0	0	n	3	2	0	0	n	n	n	n
Aulacoseira	1	2	1	0	n	2	0	1	0	0	0	0	1	0	n	3	0	0	0	n	n	n	n
Azpeitia	1	0	n	1	n	1	0	1	0	0	0	0	0	0	n	3	0	0	0	n	n	n	n
Bacillaria	2	0	n	2	0	1	0	1	1	0	0	0	0	0	n	2	0	3	0	n	n	n	n
Bacteriastrum	1	2	1	0	n	1	0	1	0	0	0	0	0	0	n	?	?	0	0	n	n	n	n
Bacterosira	1	2	1	0	n	1	0	1	0	0	0	0	0	0	n	3	4	2	0	n	n	n	n
Basilicostephanus	v	2	1	2	0	v	0	2	1	0	0	0	1	0	n	v	v	4	0	n	n	n	n
Baxteriopsis	?	1	0	0	n	1	0	?	?	0	0	0	0	0	n	?	?	0	0	n	n	n	n
Biddulphia	0	0	n	2	1	2	0	1	1	0	0	0	1	0	n	3	0	0	0	n	n	n	n
Bilingua	1	0	n	2	?	1	0	?	0	0	0	0	0	0	n	4	?	4	0	n	n	n	n
Bogorovia	0	0	1	2	0	1	0	1	1	0	0	0	0	0	n	0	n	0	0	n	n	n	n
Brightwellia	0	0	n	0	n	0	0	n	n	n	n	n	1	0	?	n	n	0	0	n	n	n	n
Caloneis	1	0	n	0	n	1	0	0	1	0	0	0	0	0	n	0	n	0	0	n	n	n	n
Campyloneis	0	0	n	0	n	0	0	n	n	n	n	n	0	0	n	n	n	0	0	n	n	n	n
Ceratoneis	0	0	n	0	n	0	0	n	n	n	n	n	0	0	n	n	n	0	0	n	n	n	n
Cereataulus	1	0	n	0	n	1	0	1	1	1	1	1	0	0	n	3	0	0	0	n	n	n	n
Cestodiscus	?	1	0	0	n	?	?	?	?	?	?	0	?	0	n	?	?	1	0	n	n	n	n
Chaetoceros	1	v	n	v	0	2	0	1	0	0	0	0	0	0	n	1	0	0	0	n	n	n	n
Charcotia	1	0	n	0	n	1	0	1	0	?	0	0	0	0	n	4	0	0	0	n	n	n	n
Cladogramma	1	0	n	0	n	1	?	?	?	?	?	?	0	0	n	?	?	2	0	n	n	n	n

Table C.2: (continued)

Genus↓ *Character* →	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Clavicula	?	0	n	0	n	?	?	?	?	?	?	0	?	0	n	?	?	0	0	n	n	n	n
Cocconeis	2	0	n	1	0	1	0	1	0	0	0	0	1	0	n	1	0	0	0	n	n	n	n
Corethron	0	0	n	0	n	2	0	1	1	1	0	0	5	2	1	2	0	0	0	n	n	n	n
Coscinodiscus	1	0	n	0	n	1	0	0	0	0	0	0	0	0	n	3	0	0	0	n	n	n	n
Cosmiodiscus	0	1	0	1	0	2	0	1	1	?	?	?	?	0	n	?	?	0	0	n	n	n	n
Craspedodiscus	1	0	n	0	n	1	0	1	0	0	0	0	0	0	n	3	2	0	0	n	n	n	n
Crucidenticula	1	0	n	0	n	2	0	1	0	0	0	0	3	0	n	2	0	0	0	n	n	n	n
Cussia	?	0	n	0	n	?	?	?	?	?	?	0	?	0	n	?	?	3	0	n	n	n	n
Cyclotella	0	0	n	0	n	1	0	1	1	1	0	0	0	0	n	2	0	5	0	n	n	n	n
Cymatodiscus	?	2	0	0	n	?	?	?	?	?	?	0	?	0	n	?	?	0	0	n	n	n	n
Cymatogonia	?	2	0	0	n	?	?	?	?	?	?	0	?	0	n	?	?	0	0	n	n	n	n
Cymatopleura	1	0	n	0	n	1	0	1	0	0	0	0	0	0	n	?	?	1	0	n	n	n	n
Cymatosira	0	2	1	0	n	1	0	1	1	0	0	0	0	0	n	2	0	0	0	n	n	n	n
Cymatotheca	?	0	n	0	n	?	1	?	?	?	?	0	?	0	n	?	?	0	0	n	n	n	n
Cymbella	1	0	n	0	n	1	1	1	1	0	0	0	0	0	n	2	0	0	0	n	n	n	n
Dactyliosolen	2	0	0	0	n	1	0	1	0	0	0	0	0	0	n	4	0	0	0	n	n	n	n
Delphineis	1	0	1	0	n	1	0	1	0	1	0	0	0	0	n	2	0	0	0	n	n	n	n
Denticula	1	0	0	0	n	1	0	1	0	0	0	0	3	0	n	2	0	0	0	n	n	n	n
Denticulopsis	2	0	n	0	n	2	0	1	1	0	0	0	0	0	n	1	0	0	0	n	n	n	n
Dextradonator	1	0	n	0	n	3	0	1	0	0	0	0	?	0	n	2	0	0	0	n	n	n	n
Diatoma	1	1	0	0	n	2	0	1	0	0	0	0	0	0	n	4	0	0	0	n	n	n	n
Dimerogramma	1	2	1	0	n	1	0	1	0	1	0	0	0	0	n	0	n	3	0	n	n	n	n
Diploneis	0	0	n	0	n	1	0	1	1	0	0	0	1	0	n	3	0	0	0	n	n	n	n
Discodiscus	0	0	n	0	n	1	0	1	1	0	0	0	1	0	n	1	2	0	1	1	0	1	0
Endictya	1	0	n	2	0	2	0	1	0	0	0	0	0	0	n	3	2	0	0	n	n	n	n
Epithemia	1	0	n	0	n	1	?	1	0	0	0	0	0	0	n	3	0	0	0	n	n	n	n
Ethmodiscus	1	0	n	1	0	1	0	1	0	0	0	0	?	0	n	3	0	0	0	n	n	n	n

Table C.2: (continued)

Genus↓ *Character* →	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Eucampia	0	0	n	0	n	2	0	1	1	0	0	0	?	0	n	?	0	0	0	n	n	n	n
Eunotia	1	0	n	0	n	2	0	1	0	0	0	0	0	0	n	4	0	3	0	n	n	n	n
Eunotogramma	1	0	0	1	0	1	0	1	0	0	0	0	?	0	n	1	0	0	0	n	n	n	n
Fenestrella	0	0	n	0	n	0	n	n	n	n	n	n	0	0	n	n	n	0	0	n	n	n	n
Fragilaria	1	2	1	0	n	2	0	1	0	1	0	0	1	0	n	2	0	0	0	n	n	n	n
Fragilariopsis	1	0	n	0	n	1	0	1	0	0	0	0	0	0	n	2	3	0	0	n	n	n	n
Gladiopsis	v	1	1	0	n	3	0	2	v	0	0	0	0	0	n	3	2	0	0	n	n	n	n
Gladius	2	0	n	2	0	3	0	2	1	0	0	0	?	0	n	3	1	1	0	n	n	n	n
Glyphodiscus	0	n	n	n	n	0	n	n	n	n	n	n	1	0	n	n	n	2	0	n	n	n	n
Gomphonema	1	0	n	0	n	1	0	1	0	0	0	0	0	0	n	2	0	0	0	n	n	n	n
Grammatophora	1	0	n	0	n	1	0	1	1	0	0	1	1	0	n	4	0	n	n	n	n	n	n
Grunowiella	1	0	n	0	n	1	0	1	0	0	0	0	?	0	n	4	0	3	0	n	n	n	n
Gyrosigma	0	0	n	0	n	1	0	1	1	0	0	0	0	0	n	3	0	0	0	n	n	n	n
Hemiaulus	0	0	n	0	n	3	1	1	1	0	0	0	0	0	n	2	4	0	0	n	n	n	n
Hemidiscus	1	0	n	0	n	1	1	1	0	0	0	0	1	0	0	3	2	0	0	n	n	n	n
Horodiscus	?	0	0	0	n	?	?	?	?	?	?	?	?	0	n	?	?	2	0	n	n	n	n
Huttonia	0	0	n	0	n	2	0	1	1	0	0	0	0	0	n	3	0	0	0	n	n	n	n
Hyalodiscus	0	0	n	0	n	0	n	n	n	n	n	n	0	1	0	n	n	0	0	n	n	n	n
Ikebea	1	1	1	0	n	1	0	?	?	?	0	0	?	0	n	0	n	0	0	n	n	n	n
Katathiraia	0	0	n	0	n	2	0	?	0	?	0	0	?	0	n	2	4	0	0	n	n	n	n
Kerkis	1	0	n	2	1	2	0	2	0	0	0	0	0	0	n	2	2	4	0	n	n	n	n
Kisseleviella	0	0	1	2	0	1	0	1	0	0	0	0	0	0	n	0	n	0	0	n	n	n	n
Kozloviella	?	1	0	?	?	?	?	?	?	?	?	?	?	?	?	?	?	5	0	n	n	n	n
Kreagra	0	0	n	0	n	3	0	1	1	0	0	0	0	0	n	4	0	4	0	n	n	n	n
Liriogramma	0	0	n	0	n	0	0	1	1	0	0	0	0	0	n	3	0	v	1	1	1	1	1
Lisitzinia	0	0	n	0	n	?	?	?	?	?	0	0	?	0	n	?	?	0	0	n	n	n	n
Lithodesmium	1	0	n	3	1	1	0	1	0	0	0	0	0	0	n	1	0	0	0	n	n	n	n

Table C.2: (continued)

Genus↓ *Character* →	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Mastogloia	0	0	n	0	n	1	0	1	1	0	0	0	1	0	n	1	0	0	0	n	n	n	n
Mediaria	1	0	n	1	0	1	1	1	0	0	0	0	0	0	n	2	1	0	n	n	n	n	0
Melosira	0	0	n	0	n	2	0	1	1	1	0	0	2	0	n	3	4	0	0	n	n	n	n
Microorbis	1	0	n	2	0	1	0	1	0	0	0	0	0	0	n	4	0	0	0	n	n	n	n
Monobrachia	0	0	n	0	n	2	0	1	0	0	0	0	?	0	n	1	4	2	0	n	n	n	n
Navicula	0	0	n	0	n	1	0	1	1	0	0	0	0	0	n	2	0	0	0	n	n	n	n
Neobrunia	0	0	n	0	n	0	0	n	n	n	n	n	0	0	n	n	n	0	0	n	n	n	n
Neodelphineis	1	1	?	0	n	1	0	1	0	0	0	0	0	0	n	2	1	3	0	n	n	n	n
Neodenticula	1	0	n	0	n	1	0	1	0	0	0	0	0	0	n	2	0	0	0	n	n	n	n
Nitzschia	0	0	n	v	?	v	1	1	1	0	0	0	v	0	n	v	0	v	0	n	n	n	n
Odontella	0	0	n	2	0	2	0	1	1	1	0	0	?	v	0	3	0	0	0	n	n	n	n
Opephora	0	0	n	0	n	1	0	1	1	0	0	0	0	0	n	2	0	0	0	n	n	n	n
Paralia	1	2	1	2	0	1	0	1	0	0	1	0	0	0	n	3	1	4	0	n	n	n	n
Peponia	0	0	n	0	n	1	0	1	1	0	0	0	1	0	n	2	0	0	0	n	n	n	n
Plagiogramma	2	1	0	0	n	1	0	1	1	1	0	0	0	0	n	2	0	0	0	n	n	n	n
Planktoniella	2	0	n	0	n	1	0	1	1	0	0	0	?	0	n	3	2	0	0	n	n	n	n
Pleurosigma	0	0	n	0	n	1	0	1	1	0	0	0	0	0	n	3	0	0	0	n	n	n	n
Podosira	0	0	n	0	n	2	0	1	1	0	0	0	1	0	n	3	2	0	0	n	n	n	n
Porosira	0	0	n	0	n	0	n	n	n	n	n	n	0	0	n	n	n	0	0	n	n	n	n
Praethalassiosiropsis	1	0	n	0	n	1	0	1	0	0	0	0	0	0	n	3	0	0	0	n	n	n	n
Pseudodimerogramma	?	0	n	0	n	?	?	?	?	?	?	?	?	0	n	?	?	0	0	n	n	n	n
Pseudoeunotia	?	0	n	0	n	?	?	?	?	?	?	?	?	?	?	?	?	0	0	n	n	n	n
Pseudopodosira	0	0	n	0	n	2	0	1	1	0	1	1	0	2	0	1	0	0	0	n	n	n	n
Pseudorutilaria	1	2	v	2	0	1	0	1	0	0	1	0	1	0	n	2	0	0	0	n	n	n	n
Pseudostictodiscus	1	0	n	0	n	1	?	?	?	?	?	0	?	0	?	?	?	0	0	n	n	n	n
Pseudotriceratium	0	0	n	0	n	0	n	n	n	n	n	n	1	0	n	n	n	0	0	n	n	n	n
Pyrgupyxis	0	0	n	0	n	1	0	?	1	?	0	0	?	0	n	1	2	0	0	n	n	n	n

Table C.2: (continued)

Genus↓ *Character* →	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Pyxilla	0	0	n	0	n	3	0	1	1	0	0	0	1	1	0	2	0	3	0	n	n	n	n
Rattrayella	1	0	n	0	n	1	0	1	0	0	0	0	0	0	n	2	0	0	0	n	n	n	n
Rhabdonema	2	0	n	0	n	2	0	1	0	0	0	0	0	1	0	0	2	0	0	n	n	n	n
Rhaphidodiscus	2	0	n	1	0	1	0	1	0	0	0	0	0	0	n	2	0	4	0	n	n	n	n
Rhaphoneis	1	0	n	2	0	1	0	1	0	0	0	0	0	0	n	2	0	0	0	n	n	n	n
Rhizosolenia	0	0	n	0	n	3	1	1	v	0	0	0	0	0	n	3	0	0	0	n	n	n	n
Rhoicosphenia	1	0	n	0	n	2	v	1	1	0	0	0	1	0	n	2	0	0	0	n	n	n	n
Rhopalodia	2	0	n	0	n	1	1	1	1	0	0	0	1	1	0	2	0	3	0	n	n	n	n
Rhynchopyxis	1	0	n	2	0	3	0	1	0	0	0	0	0	0	n	3	2	0	0	n	n	n	n
Riedelia	0	0	n	0	n	3	1	1	1	0	0	0	0	0	n	2	4	0	0	n	n	n	n
Rocella	1	0	n	0	n	1	0	1	0	0	0	0	2	0	n	2	2	0	0	n	n	n	n
Roperia	2	0	n	0	n	1	0	1	0	0	0	0	1	0	n	3	2	0	0	n	n	n	n
Rossiella	0	0	1	2	0	1	0	1	1	0	0	0	0	0	n	2	0	0	0	n	n	n	n
Rouxia	1	1	0	1	0	1	0	1	0	0	0	0	0	0	n	4	0	0	0	n	n	n	n
Rutilaria	1	1	1	0	n	1	0	1	0	0	0	0	0	0	n	0	n	0	0	n	n	n	n
Rylandsia	0	n	n	n	n	0	n	n	n	n	n	n	n	0	n	n	n	0	1	2	v	1	0
Sceptroneis	1	1	0	1	0	1	0	1	0	0	0	0	0	0	n	2	0	0	0	n	n	n	n
Simonseniella	0	1	0	0	n	3	0	1	0	0	0	0	0	0	n	0	n	0	0	n	n	n	n
Skeletonema	1	2	1	0	n	2	0	1	0	0	0	0	1	0	n	3	0	1	0	n	n	n	n
Sphynctolethus	1	0	n	2	0	2	0	1	1	0	0	0	1	1	0	2	2	0	0	n	n	n	n
Stellarima	0	n	n	n	n	0	n	n	n	n	n	n	1	0	n	n	n	0	0	n	n	n	n
Stephanodiscus	2	2	1	0	n	1	0	1	1	0	0	0	1	0	n	2	0	5	0	n	n	n	n
Stephanogonia	0	0	n	0	n	0	n	n	n	n	n	n	?	1	0	n	n	5	0	n	n	n	n
Stephanopyxis	0	2	1	0	n	2	0	1	1	0	0	0	?	0	n	3	v	0	0	n	n	n	n
Stictodiscus	1	0	n	0	n	1	0	1	0	0	0	0	1	0	n	4	0	1	0	n	n	n	n
Strangulonema	2	2	1	0	n	1	0	1	0	0	0	0	1	0	n	1	0	0	0	n	n	n	n
Surirella	1	0	n	2	1	1	0	1	0	0	0	0	1	1	0	2	0	3	0	n	n	n	n

Table C.2: (continued)

Genus↓ *Character* →	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Synedra	1	0	n	0	n	1	0	1	0	0	0	0	0	0	n	2	0	3	0	n	n	n	n
Tetracyclus	1	0	n	0	n	1	0	1	0	0	0	0	0	0	n	4	4	0	0	n	n	n	n
Thalassionema	0	0	n	0	n	1	0	1	1	0	0	0	0	0	n	0	n	0	0	n	n	n	n
Thalassiosira	2	v	0	0	n	1	0	1	0	0	0	v	2	0	n	3	v	0	0	n	n	n	n
Thalassiothrix	0	1	0	0	n	1	0	1	0	0	0	0	0	0	n	3	0	0	0	n	n	n	n
Trachyneis	0	0	n	0	n	2	0	1	1	0	0	0	1	0	n	3	0	0	0	n	n	n	n
Triceratium	1	2	?	2	0	1	0	1	0	0	1	0	1	0	n	3	0	0	0	n	n	n	n
Trinacria	1	0	n	2	0	1	0	1	0	0	0	0	1	0	n	2	0	0	0	n	n	n	n
Trochosira	1	2	0	2	0	1	0	1	0	0	0	0	0	0	n	3	0	2	0	n	n	n	n
Trochus	2	0	n	2	0	1	0	1	0	0	0	0	0	0	n	4	0	4	0	n	n	n	n
Tropidoneis	0	0	n	0	n	0	n	n	n	n	n	n	n	0	n	n	n	0	0	n	n	n	n

Table C.3: Morphospace data matrix (characters 49-71)

Genus↓ *Character* →	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71
Abas	0	0	0	n	2	0	0	0	0	0	0	0	0	0	n	n	n	n	0	1	0	0	0
Achnanthes	0	0	0	n	0	0	0	0	0	0	0	1	2	n	0	n	n	v	0	1	0	1	0
Actinocyclus	0	1	0	n	0	0	0	0	0	0	0	0	2	0	n	n	n	0	0	0	0	2	0
Actinoptychus	v	1	1	1	0	0	0	0	0	0	0	0	2	n	0	n	n	0	0	0	0	1	0
Amblypyrgus	0	0	0	n	0	0	0	0	0	0	2	0	0	0	n	n	n	n	0	2	0	0	0
Amphora	0	0	0	n	0	0	0	0	0	0	0	1	2	n	1	0	1	v	0	v	v	v	0
Anaulus	0	0	0	n	0	0	1	0	0	0	0	0	0	0	n	n	n	n	0	1	0	1	0
Ancylopyrgus	0	0	0	n	0	0	0	0	0	0	2	0	0	4	n	n	n	n	0	2	0	0	0
Annellus	0	0	1	0	0	0	0	0	0	0	0	0	0	0	n	n	n	n	0	2	0	0	0
Arachnoidiscus	0	0	1	v	0	0	0	0	0	0	0	0	2	n	0	n	n	0	0	1	0	0	0
Archepyrgus	0	0	1	1	0	0	0	0	0	0	0	0	0	0	n	n	n	n	0	0	0	1	0
Asterolampra	0	0	0	n	0	0	0	0	0	0	0	0	0	0	n	n	n	n	0	1	0	0	0

Table C.3: (continued)

Genus↓ *Character* →	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71
Asteromphalus	0	0	0	n	0	0	0	0	0	0	0	0	0	0	n	n	n	n	0	1	0	0	0
Aulacodiscus	0	1	1	1	0	0	0	0	0	0	0	0	2	n	0	n	n	0	1	1	0	1	3
Aulacoseira	0	0	0	n	0	0	0	0	0	0	0	0	0	4	n	n	n	n	0	1	0	1	0
Azpeitia	1	0	0	n	0	0	0	0	0	0	0	0	0	2	n	n	n	n	1	1	0	0	0
Bacillaria	0	0	0	n	0	0	0	0	0	0	0	1	2	n	0	n	n	0	0	0	0	2	0
Bacteriastrum	0	0	0	n	0	2	0	0	0	0	0	0	?	?	?	?	?	?	0	0	?	?	0
Bacterosira	0	0	0	n	0	0	0	0	0	0	0	0	2	n	0	n	n	0	0	1	0	1	0
Basilicostephanus	0	0	0	n	0	0	0	0	0	0	0	0	1	n	0	n	n	0	0	1	0	1	0
Baxteriopsis	0	0	0	n	0	0	0	0	1	0	0	0	3	n	n	n	n	n	0	1	0	0	3
Biddulphia	0	0	0	n	0	0	0	0	0	0	0	0	0	0	n	n	n	n	2	2	0	0	0
Bilingua	0	0	0	n	0	0	0	0	0	0	2	0	3	n	0	n	n	0	0	0	0	1	0
Bogorovia	0	0	0	n	0	0	0	0	0	0	0	?	3	n	n	n	n	n	0	v	0	v	0
Brightwellia	0	0	1	1	0	0	0	0	0	0	0	0	0	3	n	n	n	n	0	1	0	2	0
Caloneis	1	0	0	n	0	0	0	0	0	0	0	1	2	n	0	n	n	2	0	0	0	1	0
Campyloneis	0	0	0	n	0	0	0	0	0	0	0	1	2	n	1	0	1	0	0	0	0	2	0
Ceratoneis	0	1	0	n	0	0	0	0	0	0	0	1	?	?	?	?	?	?	?	0	?	?	?
Cereataulus	1	0	0	n	0	0	0	0	0	0	0	0	0	3	n	n	n	n	0	1	0	0	0
Cestodiscus	0	0	0	n	0	0	0	0	0	0	0	0	v	1	0	n	n	0	1	2	0	0	3
Chaetoceros	0	0	0	n	0	1	0	0	0	0	0	0	3	n	n	n	n	n	0	0	0	2	0
Charcotia	0	0	1	2	0	0	0	0	0	0	0	0	2	n	0	n	n	0	0	1	0	2	0
Cladogramma	?	1	0	n	0	0	0	0	0	0	0	?	n	n	n	n	n	1	0	2	0	2	0
Clavicula	0	0	0	n	0	0	0	0	0	0	0	1	3	n	n	n	n	n	0	2	0	2	1
Cocconeis	0	0	0	n	0	0	0	0	0	0	0	1	2	n	1	0	1	0	0	1	0	2	0
Corethron	1	0	0	n	0	0	0	0	1	0	0	0	2	n	0	n	n	0	0	0	0	1	0
Coscinodiscus	0	0	0	n	0	0	0	0	0	0	0	0	0	4	n	n	n	n	1	2	0	1	0
Cosmiodiscus	0	0	1	1	0	0	0	0	0	0	0	0	2	n	0	n	n	0	1	2	0	1	0
Craspedodiscus	1	0	1	1	0	0	0	0	0	0	0	0	0	3	n	n	n	n	2	2	0	0	0

Table C.3: (continued)

Genus↓ *Character* →	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71
Crucidenticula	0	0	0	n	0	0	0	0	0	0	0	1	2	n	0	n	n	0	0	2	0	2	0
Cussia	?	?	0	n	0	0	0	?	?	0	0	?	3	n	n	n	n	n	0	v	0	1	0
Cyclotella	1	0	1	v	0	0	0	0	0	0	0	0	2	0	0	n	n	v	0	0	0	2	0
Cymatodiscus	0	0	1	2	0	0	0	0	0	0	0	0	2	n	0	n	n	0	1	1	0	1	1
Cymatogonia	0	0	1	1	0	0	0	?	?	0	0	0	0	0	n	n	n	n	1	1	?	0	0
Cymatopleura	0	0	0	n	0	0	0	0	0	0	0	1	2	n	0	n	n	0	0	0	0	2	0
Cymatosira	0	0	0	n	0	0	0	0	0	0	0	?	2	n	?	n	n	0	0	2	0	2	0
Cymatotheca	2	0	0	n	0	0	0	0	0	0	0	0	2	n	0	n	n	0	1	v	0	0	1
Cymbella	0	0	0	n	0	0	0	0	0	0	0	1	2	n	0	n	n	0	0	1	5	1	0
Dactyliosolen	0	0	0	n	0	0	0	0	0	0	0	0	3	n	n	n	n	n	0	0	0	2	0
Delphineis	0	0	0	n	0	0	0	0	0	0	0	1	2	n	1	0	1	0	0	2	0	2	0
Denticula	0	0	0	n	0	0	0	0	0	0	0	1	2	n	1	0	1	0	0	1	0	2	0
Denticulopsis	0	0	0	n	0	0	0	0	0	0	0	1	2	n	0	n	n	2	0	0	0	2	0
Dextradonator	0	0	0	n	0	0	0	0	0	0	1	0	3	n	n	n	n	n	0	1	0	2	0
Diatoma	0	0	0	n	0	0	0	0	0	0	0	1	2	n	0	n	n	0	0	0	0	2	0
Dimerogramma	0	0	0	n	0	0	0	0	0	0	0	1	2	n	1	0	1	0	0	2	0	1	0
Diploneis	0	0	0	n	0	0	0	0	0	0	0	1	2	n	1	0	1	0	v	2	v	2	0
Discodiscus	0	0	0	n	0	0	0	0	0	0	0	0	0	4	n	n	n	n	1	1	0	0	0
Endictya	1	1	0	n	2	0	0	0	0	0	0	0	0	3	n	n	n	n	3	2	0	1	0
Epithemia	1	0	0	n	0	0	0	0	0	0	0	1	2	n	0	n	n	0	0	1	v	0	0
Ethmodiscus	1	0	1	1	0	0	0	0	0	0	0	0	2	n	0	n	n	0	0	0	0	2	0
Eucampia	0	0	0	n	0	0	0	0	0	0	1	0	2	n	0	n	n	0	0	2	4	0	0
Eunotia	0	0	0	n	0	0	0	0	0	0	0	1	2	n	1	0	1	0	0	0	0	2	0
Eunotogramma	0	0	0	n	0	0	0	0	0	0	0	0	2	0	0	n	n	0	0	1	0	1	0
Fenestrella	0	0	0	n	0	0	0	0	0	0	0	0	0	1	n	n	n	n	0	0	0	1	0
Fragilaria	0	0	0	n	0	0	0	0	0	0	0	1	2	n	0	n	n	0	0	0	0	1	0
Fragilariopsis	0	0	0	n	0	0	0	0	0	0	0	1	2	n	0	n	n	1	0	0	0	1	0

Table C.3: (continued)

Genus↓ *Character* →	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71
Gladiopsis	0	0	1	0	0	0	0	0	0	0	0	0	3	0	0	n	n	0	0	0	0	2	0
Gladius	0	0	0	n	0	0	0	0	0	0	2	0	0	0	n	n	n	n	0	0	0	1	0
Glyphodiscus	1	3	0	n	1	0	0	0	0	0	0	0	2	n	0	n	n	0	0	1	0	1	0
Gomphonema	0	0	0	n	0	0	0	0	0	0	0	1	2	n	1	v	0	0	0	1	0	2	0
Grammatophora	0	0	0	n	0	0	0	0	0	0	0	1	2	n	0	n	n	0	0	0	0	1	0
Grunowiella	0	0	0	n	0	0	0	0	0	0	0	1	1	n	0	n	n	0	0	1	1	2	0
Gyrosigma	0	0	0	n	0	0	0	0	0	0	0	1	1	n	0	n	n	n	0	1	5	1	0
Hemiaulus	0	0	0	n	0	0	0	0	0	0	1	0	v	v	n	n	n	n	v	v	v	v	0
Hemidiscus	0	0	0	n	0	0	0	0	0	0	0	0	0	1	n	n	n	n	0	1	0	1	3
Horodiscus	1	3	1	0	0	0	0	0	0	0	0	0	?	?	?	?	?	?	?	?	?	?	?
Huttonia	0	0	0	n	0	0	0	0	0	0	0	0	3	n	n	n	n	n	0	1	0	0	0
Hyalodiscus	0	0	1	1	0	0	0	0	0	0	0	0	0	1	n	n	n	n	0	0	0	1	0
Ikebea	0	0	0	n	0	0	0	0	0	0	0	1	2	n	0	n	n	0	0	v	0	2	0
Katathiraia	?	0	0	n	0	0	0	0	0	0	0	1	3	n	n	n	n	n	1	2	1	0	0
Kerkis	0	0	0	n	0	0	0	0	0	0	1	0	2	n	0	n	n	0	0	0	0	1	?
Kisseleviella	0	0	1	1	2	0	0	0	0	0	0	0	3	n	n	n	n	n	0	0	0	2	0
Kozloviella	?	0	1	1	0	0	0	0	0	0	0	0	2	n	0	n	n	0	1	1	0	1	3
Kreagra	0	0	0	n	0	0	0	0	0	0	2	0	2	n	v	n	n	0	0	0	0	?	0
Liriogramma	0	0	0	n	0	0	0	0	0	0	0	0	0	0	n	n	n	n	0	1	0	0	0
Lisitzinia	0	0	1	2	0	0	0	0	0	0	0	0	0	0	n	n	n	n	0	2	0	0	0
Lithodesmium	0	0	0	n	0	0	1	0	0	1	0	0	2	n	0	n	n	1	0	1	0	1	0
Mastogloia	0	0	0	n	0	0	0	0	0	0	0	1	v	v	v	v	v	v	v	1	0	v	v
Mediaria	0	0	0	n	0	0	0	0	0	0	0	1	2	n	2	0	2	0	0	1	0	0	0
Melosira	1	0	0	n	4	0	0	0	0	1	0	0	3	n	n	n	n	n	0	0	0	2	1
Microorbis	1	0	1	0	0	0	0	0	0	0	0	0	2	n	0	n	n	0	0	0	0	1	0
Monobrachia	0	0	0	n	0	0	0	0	0	0	0	0	3	n	n	n	n	0	0	0	2	0	?
Navicula	0	0	0	n	0	0	0	0	0	0	0	1	1	n	1	1	v	0	0	0	5	0	0

Table C.3: (continued)

Genus↓ *Character* →	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71
Neobrunia	0	0	1	v	0	0	0	0	0	0	0	0	0	1	n	n	n	n	1	1	v	0	?
Neodelphineis	0	0	0	n	0	0	0	0	0	0	0	1	1	n	1	0	1	0	0	1	0	2	0
Neodenticula	0	0	0	n	0	0	0	0	0	0	0	1	2	n	2	n	n	2	0	0	0	0	0
Nitzschia	0	0	0	n	0	0	0	0	0	0	0	1	1	n	2	n	n	0	0	1	0	1	0
Odontella	1	0	0	n	0	0	1	0	0	0	0	0	0	?	n	n	n	n	0	0	0	2	0
Opephora	0	0	0	n	n	0	0	0	0	0	0	1	2	n	0	n	n	0	0	3	4	0	0
Paralia	0	0	1	1	0	0	0	0	0	0	0	0	2	n	0	n	n	0	0	2	0	2	0
Peponia	n	0	0	n	1	0	0	0	0	0	0	0	0	0	n	n	n	n	1	1	0	1	0
Plagiogramma	0	0	0	n	0	0	0	0	0	0	0	1	1	n	0	n	n	0	0	1	1	1	0
Planktoniella	0	0	0	n	0	0	0	0	0	0	0	0	0	0	n	n	n	n	0	1	0	1	0
Pleurosigma	0	0	0	n	0	0	0	0	0	0	0	1	0	0	n	n	n	n	0	0	5	1	0
Podosira	0	0	0	n	0	0	0	0	0	0	0	0	0	1	n	n	n	n	0	0	0	0	0
Porosira	0	0	0	n	0	0	0	0	0	0	0	0	0	4	n	n	n	n	0	0	5	1	0
Praethalassiosiropsis	0	0	1	0	0	0	0	0	0	0	0	0	0	4	n	n	n	n	1	2	0	0	0
Pseudodimerogramma	?	0	0	n	n	?	?	?	?	0	0	1	2	n	v	v	v	0	0	1	0	1	0
Pseudoeunotia	?	?	0	n	0	0	0	?	?	0	0	1	2	n	2	n	n	1	0	?	?	?	0
Pseudopodosira	0	0	1	1	1	0	0	0	0	0	0	0	0	0	n	n	n	n	0	0	0	1	0
Pseudorutilaria	0	1	1	2	0	0	2	0	0	0	1	0	3	n	n	n	n	n	0	1	0	2	0
Pseudostictodiscus	0	0	1	1	0	0	0	?	?	0	0	0	2	n	0	n	n	0	2	1	0	0	0
Pseudotriceratium	0	0	0	n	0	0	0	0	0	0	0	0	0	2	n	n	n	n	1	1	0	1	4
Pyrgupyxis	?	0	0	n	0	0	0	0	0	0	2	0	0	0	n	n	n	n	2	1	0	0	3
Pyxilla	0	1	0	n	0	0	0	0	0	0	2	0	0	0	n	n	n	n	0	2	0	0	0
Rattrayella	0	0	0	n	0	0	0	0	0	0	0	0	2	n	0	n	n	0	0	0	0	2	0
Rhabdonema	0	0	0	n	0	0	0	0	0	0	0	1	2	n	0	n	n	0	0	1	0	1	0
Rhaphidodiscus	0	0	2	0	0	0	0	0	0	0	0	1	1	n	1	1	1	0	1	1	0	1	3
Rhaphoneis	1	0	0	n	0	0	0	0	0	0	0	1	2	n	1	0	1	0	1	2	0	1	0
Rhizosolenia	0	0	0	n	0	0	0	0	0	0	2	0	2	n	0	n	n	0	0	0	0	0	0

Table C.3: (continued)

Genus↓ *Character* →	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71
Rhoicosphenia	0	0	0	n	0	0	0	0	0	0	0	1	1	n	0	n	n	0	0	0	5	1	0
Rhopalodia	0	0	0	n	0	0	0	0	0	0	0	1	2	n	0	n	n	v	0	1	0	0	0
Rhynchopyxis	0	0	0	n	0	0	1	0	0	0	2	0	0	4	n	n	n	n	0	2	0	0	0
Riedelia	0	0	0	n	0	0	0	0	0	0	1	0	v	v	n	n	n	n	v	v	v	v	0
Rocella	0	0	0	n	0	0	0	0	0	0	0	0	0	n	n	n	n	n	1	3	0	0	0
Roperia	0	0	0	n	0	0	0	0	0	0	0	0	0	0	n	n	n	n	1	1	0	0	0
Rossiella	0	0	0	n	0	0	0	0	0	0	0	0	0	v	n	n	n	n	0	2	0	0	0
Rouxia	0	0	0	n	0	0	0	0	0	0	2	1	2	n	0	n	n	0	0	1	3	0	0
Rutilaria	0	0	0	n	n	0	1	0	0	0	1	0	2	n	0	n	n	0	0	0	0	1	0
Rylandsia	0	0	0	n	0	0	0	0	0	0	0	0	0	0	n	n	n	n	1	1	0	0	0
Sceptroneis	0	0	0	n	0	0	0	0	0	0	0	1	1	n	0	n	n	0	0	2	0	1	0
Simonseniella	0	0	0	n	0	0	0	0	0	0	0	?	2	n	?	?	?	0	0	0	0	2	?
Skeletonema	0	0	0	n	0	0	0	0	0	0	0	0	0	1	n	n	n	n	0	1	4	0	0
Sphynctolethus	0	0	0	n	0	0	0	0	0	0	1	0	2	n	n	0	n	n	2	1	1	1	0
Stellarima	0	0	0	n	0	0	0	0	0	0	0	0	0	1	n	n	n	n	0	0	0	1	0
Stephanodiscus	0	0	0	n	0	0	0	0	0	0	0	0	2	n	0	n	n	v	0	0	0	2	0
Stephanogonia	0	1	1	0	2	0	0	0	0	0	0	0	3	n	n	n	n	n	0	0	0	2	0
Stephanopyxis	0	0	0	n	0	0	0	0	0	0	v	0	0	4	n	n	n	n	1	2	0	0	0
Stictodiscus	0	0	1	2	0	0	0	0	0	0	0	0	0	3	n	n	n	n	0	0	0	2	0
Strangulonema	0	0	0	n	0	0	0	0	0	1	0	0	0	1	n	n	n	n	0	0	1	0	0
Surirella	1	1	0	n	0	0	0	0	0	0	0	1	2	n	1	0	1	2	0	0	0	1	0
Synedra	0	0	0	n	0	0	0	0	0	0	0	1	2	n	0	n	n	0	0	0	0	2	0
Tetracyclus	0	0	0	n	0	0	0	0	0	0	0	1	2	n	1	1	1	0	0	0	2	0	0
Thalassionema	0	0	0	n	0	0	0	0	0	0	0	1	1	n	0	n	n	0	0	2	0	0	0
Thalassiosira	0	0	0	n	0	0	0	0	0	0	0	0	0	v	n	n	n	n	1	1	0	1	0
Thalassiothrix	0	0	0	n	0	0	0	0	0	0	0	1	2	n	0	n	n	0	0	2	3	0	0
Trachyneis	0	0	0	n	0	0	0	0	0	0	0	1	2	n	1	1	1	0	0	5	1	0	0

Table C.3: (continued)

Genus↓ *Character* →	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71
Triceratium	0	0	n	0	0	0	0	1	0	0	0	0	0	0	n	n	n	n	0	2	0	0	0
Trinacria	0	0	0	n	0	0	0	1	0	0	1	0	2	n	0	n	n	0	1	2	0	1	0
Trochosira	0	1	1	1	0	0	0	0	0	0	2	0	2	n	0	n	n	0	0	0	0	1	0
Trochus	0	0	1	2	0	0	0	0	0	0	0	0	1	n	0	n	n	0	0	1	0	1	3
Tropidoneis	0	0	0	n	0	0	0	0	0	0	0	1	1	n	0	n	n	0	0	0	0	1	0

Table C.4: Morphospace data matrix (characters 72-94)

Genus↓ *Character* →	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
Abas	?	?	0	?	?	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Achnanthes	2	0	0	?	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	v	0	0	n
Actinocyclus	0	2	?	2	0	0	0	0	1	1	0	0	0	n	0	0	n	0	0	n	n	n	n
Actinoptychus	0	2	?	2	0	0	0	0	1	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Amblypyrgus	1	2	0	0	1	0	0	0	0	0	0	0	0	n	1	0	n	0	0	n	n	n	n
Amphora	v	2	v	?	0	0	0	0	0	0	0	v	0	0	0	?	?	?	?	0	n	n	n
Anaulus	0	2	?	?	0	0	0	0	0	0	0	0	0	?	?	1	?	?	?	1	?	0	n
Ancylopyrgus	1	2	0	0	1	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Annellus	0	2	?	?	0	0	0	0	0	0	0	0	0	n	1	0	n	0	0	n	n	n	n
Arachnoidiscus	2	0	2	1	0	0	1	0	0	0	1	1	0	0	1	1	0	0	1	0	n	n	n
Archepyrgus	?	?	?	?	0	0	0	0	0	0	0	0	0	n	1	0	n	0	0	n	n	n	n
Asterolampra	2	1	0	0	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Asteromphalus	2	1	0	0	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Aulacodiscus	2	1	1	0	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Aulacoseira	0	2	?	?	0	0	0	0	0	0	0	0	0	n	1	0	n	0	0	n	n	n	n
Azpeitia	2	1	0	1	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Bacillaria	1	2	?	1	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Bacteriastrum	?	?	?	?	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n

Table C.4: (continued)

Genus↓ *Character* →	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
Bacterosira	1	2	0	0	1	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Basilicostephanus	0	0	n	n	0	0	0	0	0	0	0	0	0	n	1	0	n	0	0	n	n	n	n
Baxteriopsis	?	?	?	?	0	?	?	?	?	0	0	0	0	?	?	?	?	?	?	2	?	0	n
Biddulphia	2	1	0	1	0	0	0	1	0	0	0	0	0	0	1	2	0	0	0	2	0	0	n
Bilingua	0	0	n	n	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Bogorovia	?	?	?	?	0	0	0	0	0	0	0	0	0	0	0	0	n	1	0	1	?	0	n
Brightwellia	2	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	n	0	0	n	n	n	n
Caloneis	0	0	n	n	0	1	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Campyloneis	0	0	n	n	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Ceratoneis	?	?	?	?	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Cereataulus	2	1	0	0	0	0	0	0	0	0	0	1	0	0	?	0	n	0	0	1	0	0	n
Cestodiscus	?	?	?	?	0	0	?	0	0	0	0	1	0	0	0	0	n	0	0	n	n	n	n
Chaetoceros	0	0	n	n	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	1	0	0	n
Charcotia	2	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	n	0	0	0	n	n	n
Cladogramma	?	?	?	?	0	0	?	?	?	?	0	0	1	?	?	?	?	?	?	n	n	n	n
Clavicula	?	?	?	?	0	0	?	?	?	0	0	0	1	?	?	?	?	?	?	2	?	n	n
Cocconeis	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	n	0	0	0	n	n	n
Corethron	?	?	?	?	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	n	n	n	n
Coscinodiscus	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	n	n	n	n
Cosmiodiscus	?	?	?	?	0	0	?	?	?	?	0	1	0	?	?	?	?	?	?	n	n	n	n
Craspedodiscus	2	1	0	1	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Crucidenticula	2	0	1	1	0	0	0	0	0	0	0	1	0	0	1	2	1	1	1	0	n	n	n
Cussia	?	?	?	?	0	0	?	?	?	0	0	0	0	?	?	?	?	?	?	0	n	n	n
Cyclotella	1	2	0	2	0	1	0	0	0	0	0	0	0	0	0	0	n	0	0	n	n	n	n
Cymatodiscus	?	?	?	?	0	0	?	?	?	?	0	0	0	?	?	0	n	0	0	0	n	n	n
Cymatogonia	?	?	?	?	0	0	?	?	?	?	0	0	0	?	?	0	n	0	0	0	n	n	n
Cymatopleura	0	2	?	?	0	0	0	0	0	0	0	0	0	0	1	0	n	0	1	0	n	n	n

Table C.4: (continued)

Genus↓ *Character* →	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
Cymatosira	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	3	1	0	n
Cymatotheca	?	?	?	?	0	?	?	?	?	?	0	0	1	0	0	0	n	0	0	0	n	n	n
Cymbella	0	2	2	?	0	0	0	0	0	0	0	0	0	0	0	0	j	2	0	2	2	0	n
Dactyliosolen	0	0	n	n	0	0	0	0	0	0	0	0	0	n	0	0	n	2	0	n	n	n	n
Delphineis	2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Denticula	0	2	0	1	0	0	0	0	0	0	0	0	0	0	1	2	1	0	0	0	n	n	n
Denticulopsis	0	0	n	n	0	0	0	0	0	0	0	0	0	1	1	2	1	2	1	0	n	n	n
Dextradonator	?	?	?	?	0	0	?	0	0	0	0	0	0	n	0	0	n	0	0	0	n	n	n
Diatoma	0	0	n	n	0	0	0	0	0	0	0	0	0	0	1	0	n	1	0	2	0	0	n
Dimerogramma	2	?	0	1	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	2	1	0	n
Diploneis	2	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	n	0	0	0	n	n	n
Discodiscus	?	?	?	?	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Endictya	1	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	n	0	0	n	n	n	n
Epithemia	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	n	n	n
Ethmodiscus	2	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	n	n	n	n
Eucampia	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	n	?	0	1	1	1	n
Eunotia	0	0	n	n	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Eunotogramma	?	?	?	?	?	0	0	0	0	0	0	0	0	0	0	2	?	1	0	0	n	n	n
Fenestrella	2	1	0	0	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Fragilaria	2	0	v	1	0	0	0	0	0	0	0	0	0	0	0	0	n	2	0	4	1	0	n
Fragilariopsis	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	n	2	0	0	n	n	n
Gladiopsis	0	0	n	n	0	0	0	0	0	0	0	0	0	n	1	0	n	0	0	0	n	n	n
Gladius	0	0	n	n	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Glyphodiscus	1	0	n	n	0	0	0	0	0	0	0	0	1	0	0	0	n	0	0	1	0	0	n
Gomphonema	2	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	n	2	0	0	1	n	2
Grammatophora	0	0	n	n	0	0	0	0	0	0	0	0	0	1	2	0	n	0	0	2	0	0	n
Grunowiella	0	0	n	n	0	0	0	0	0	0	0	0	0	?	?	?	?	?	?	2	?	?	0

Table C.4: (continued)

Genus↓ *Character* →	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
Gyrosigma	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Hemiaulus	2	0	0	?	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Hemidiscus	2	1	0	1	0	0	0	0	1	1	0	0	0	0	1	0	n	0	0	0	n	n	n
Horodiscus	?	?	?	?	?	?	?	?	?	0	0	0	0	n	?	?	?	?	?	n	n	n	n
Huttonia	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	n
Hyalodiscus	0	2	1	?	0	0	0	0	1	0	0	0	0	0	0	0	n	0	0	n	n	n	n
Ikebea	?	?	?	?	0	0	0	0	0	0	0	0	0	?	?	0	n	0	?	1	?	?	1
Katathiraia	2	0	0	?	0	0	0	0	0	0	0	0	0	?	?	2	0	1	1	0	n	n	n
Kerkis	0	0	n	n	0	0	0	0	0	0	0	0	0	1	1	0	n	0	0	0	n	n	n
Kisseleviella	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	3	0	0	n
Kozloviella	2	?	0	?	0	0	0	0	0	0	0	0	0	?	?	0	n	0	0	0	n	n	n
Kreagra	0	0	n	n	0	0	0	0	0	0	0	1	0	n	0	0	n	0	0	n	n	n	n
Liriogramma	2	1	0	0	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Lisitzinia	?	?	?	?	0	0	?	?	0	0	0	0	0	0	1	2	?	0	0	?	?	?	n
Lithodesmium	0	0	n	n	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	?	?	?	n
Mastogloia	v	v	v	v	0	0	0	0	0	0	0	0	0	1	2	0	n	2	0	0	n	n	n
Mediaria	2	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Melosira	1	2	1	?	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Microorbis	0	0	n	n	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Monobrachia	?	?	?	?	0	0	?	0	0	0	0	0	0	?	?	?	?	?	?	0	n	n	n
Navicula	0	2	0	?	0	0	0	0	0	0	0	0	0	0	0	0	n	2	0	0	n	n	n
Neobrunia	2	2	2	1	0	0	0	0	0	0	2	0	0	n	0	0	n	0	0	n	n	n	n
Neodelphineis	2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Neodenticula	0	0	n	n	0	0	0	0	0	0	0	0	0	0	1	2	1	1	1	0	n	n	n
Nitzschia	v	v	v	v	0	0	0	0	0	0	0	0	0	0	0	0	n	v	0	0	n	n	n
Odontella	2	1	0	0	0	0	0	0	0	0	0	0	0	?	?	?	?	?	?	1	0	0	n
Opephora	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	2	1	1	2

Table C.4: (continued)

Genus↓ *Character* →	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
Paralia	1	1	n	n	0	1	0	1	0	0	0	0	1	0	0	0	n	2	0	n	n	n	n
Peponia	0	0	n	n	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	1	?	?	n
Plagiogramma	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	2	0	2	0	0	n
Planktoniella	1	2	0	0	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Pleurosigma	1	2	0	?	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Podosira	1	2	1	0	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Porosira	1	2	0	0	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Praethalassiosiropsis	1	2	0	0	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Pseudodimerogramma	?	?	?	?	0	0	?	?	?	0	0	0	0	?	?	0	n	0	0	0	n	n	4
Pseudoeunotia	?	?	?	?	?	?	?	?	?	?	0	0	0	?	?	?	?	2	?	?	?	?	?
Pseudopodosira	?	?	?	?	0	0	0	0	0	0	0	1	1	n	0	0	n	0	0	n	n	n	n
Pseudorutilaria	0	2	2	1	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	1	0	0	n
Pseudostictodiscus	?	?	?	?	0	0	?	?	?	0	0	0	1	n	0	0	n	0	0	n	n	n	n
Pseudotriceratium	1	1	n	n	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Pyrgupyxis	?	?	?	?	0	0	?	?	?	0	0	0	0	n	?	?	?	?	?	n	n	n	n
Pyxilla	2	0	0	1	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Rattrayella	0	0	n	n	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	2	0	0	n
Rhabdonema	2	0	v	0	0	0	0	0	0	0	0	0	0	1	2	0	n	2	0	2	1	0	n
Rhaphidodiscus	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	n	2	0	0	n	n	n
Rhaphoneis	2	0	1	0	0	0	0	0	0	0	0	0	0	1	2	0	n	0	0	2	0	0	n
Rhizosolenia	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	n	n	n	n
Rhoicosphenia	0	2	0	?	0	0	0	0	0	0	0	0	0	1	2	0	n	2	0	2	1	0	2
Rhopalodia	0	2	2	0	?	0	0	0	0	0	0	0	0	0	0	0	n	2	0	0	n	n	n
Rhynchopyxis	1	2	0	0	1	0	0	0	0	0	0	0	0	n	1	0	n	0	0	n	n	n	n
Riedelia	0	0	n	n	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Rocella	0	2	?	?	0	0	0	0	0	0	0	0	0	n	1	0	n	0	1	n	n	n	n
Roperia	2	1	0	0	0	0	0	0	0	1	0	0	0	n	0	0	n	0	0	n	n	n	n

Table C.4: (continued)

Genus↓ *Character* →	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
Rossiella	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	1	?	0	n
Rouxia	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Rutilaria	?	?	?	?	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	1	?	0	n
Rylandsia	2	1	0	0	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Sceptroneis	2	0	1	?	0	0	0	0	0	0	0	0	0	0	0	0	n	0	1	2	0	0	2
Simonseniella	?	?	?	?	0	0	0	0	0	0	0	1	1	0	0	0	n	0	0	n	n	n	n
Skeletonema	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	n	0	0	n	n	n	n
Sphynctolethus	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	n	0	0	1	1	1	n
Stellarima	2	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	n	n	n	n
Stephanodiscus	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	n	n	n	n
Stephanogonia	0	0	n	n	0	0	0	0	0	0	0	0	0	?	?	?	?	?	?	n	n	n	n
Stephanopyxis	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	n	n	n	n
Stictodiscus	0	0	n	n	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Strangulonema	0	0	n	n	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Surirella	0	2	2	?	0	0	0	0	0	0	0	0	0	0	0	0	n	2	1	0	n	n	0
Synedra	?	?	?	?	0	0	0	0	0	0	0	0	0	0	0	0	n	2	1	4	1	0	n
Tetracyclus	0	0	n	n	0	0	0	0	0	0	0	0	0	1	1	0	n	1	1	0	n	n	n
Thalassionema	2	1	3	2	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Thalassiosira	1	2	0	2	0	0	0	0	0	0	0	0	0	n	1	0	n	0	0	n	n	n	n
Thalassiothrix	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	0
Trachyneis	1	2	0	2	0	1	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n
Triceratium	1	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	1	0	0	n
Trinacria	2	0	0	v	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	2	?	?	n
Trochosira	2	1	0	?	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Trochus	0	0	n	n	0	0	0	0	0	0	0	0	0	n	0	0	n	0	0	n	n	n	n
Tropidoneis	2	1	0	?	0	0	0	0	0	0	0	0	0	0	0	0	n	0	0	0	n	n	n

Table C.5: Morphospace data matrix (characters 95-113)

Genus↓ *Character* →	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113
Abas	n	n	2	0	n	n	1	2	3	?	?	0	n	n	n	n	n	0	n
Achnanthes	n	0	n	0	n	n	0	n	n	n	0	0	0	0	0	0	1	1	1
Actinocyclus	n	n	n	0	n	n	3	2	0	2	3	0	n	n	n	n	0	0	n
Actinoptychus	n	0	0	0	n	n	3	2	1	0	0	0	n	n	n	n	n	0	n
Amblypyrgus	n	n	n	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Amphora	n	0	n	0	n	n	0	n	n	n	n	0	0	0	1	0	1	2	1
Anaulus	n	0	0	0	n	n	1	0	1	1	0	0	n	n	n	n	n	n	n
Ancylopyrgus	n	n	n	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Annellus	n	n	n	0	n	n	0	n	n	n	n	0	n	n	n	n	0	0	n
Arachnoidiscus	n	n	n	0	n	n	3	0	?	?	?	0	n	n	n	n	n	n	n
Archepyrgus	n	n	n	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Asterolampra	n	n	n	0	n	n	3	2	0	2	2	0	n	n	n	n	n	n	n
Asteromphalus	n	n	n	0	n	n	3	2	0	2	2	0	n	n	n	n	n	n	n
Aulacodiscus	n	n	n	0	n	n	3	2	3	0	0	0	n	n	n	n	n	n	n
Aulacoseira	n	n	n	0	n	n	3	3	0	0	0	0	n	n	n	n	n	n	n
Azpeitia	n	n	n	0	n	n	3	3	0	1	?	0	n	n	n	n	n	n	n
Bacillaria	n	0	n	0	n	n	0	n	n	n	n	0	0	0	0	0	0	1	0
Bacteriastrum	n	n	n	0	n	n	1	0	0	?	?	0	n	n	n	n	n	n	n
Bacterosira	n	n	n	3	0	0	1	2	0	0	0	0	n	n	n	n	n	n	n
Basilicostephanus	n	n	n	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Baxteriopsis	n	2	?	?	?	?	?	?	?	?	?	?	n	n	n	n	n	n	n
Biddulphia	n	0	0	0	n	n	3	0	3	0	0	0	n	n	n	n	n	n	n
Bilingua	n	0	0	0	n	n	3	0	0	?	?	0	n	n	n	n	n	n	n
Bogorovia	n	0	n	0	n	n	1	2	0	0	0	0	n	n	n	n	n	n	n
Brightwellia	n	n	n	0	n	n	3	2	0	?	?	0	n	n	n	n	n	n	n
Caloneis	n	0	n	0	n	n	0	n	n	n	n	0	0	1	0	0	0	1	0
Campyloneis	n	0	n	0	n	n	n	n	n	n	n	0	0	2	0	0	0	1	1

Table C.5: (continued)

Genus↓ *Character* →	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113
Ceratoneis	n	n	n	0	n	n	0	n	n	n	n	0	0	2	1	0	0	2	v
Cereataulus	n	n	n	0	n	n	1	2	1	?	?	0	n	n	n	n	n	n	n
Cestodiscus	n	n	n	0	n	n	3	2	?	?	?	0	n	n	n	n	n	n	n
Chaetoceros	n	0	0	0	n	n	1	0	?	?	?	0	n	n	n	n	n	n	n
Charcotia	n	n	n	0	n	n	3	2	0	1	0	0	n	n	n	n	n	n	n
Cladogramma	n	n	n	?	?	?	?	?	?	?	?	?	n	n	n	n	n	n	n
Clavicula	0	?	?	?	?	?	?	?	?	?	?	?	?	2	?	?	?	0	n
Cocconeis	n	n	n	0	n	n	0	n	n	n	n	0	0	0	0	0	0	1	1
Corethron	n	n	n	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Coscinodiscus	n	n	n	0	n	n	3	2	0	0	0	1	n	n	n	n	n	n	n
Cosmiodiscus	n	n	n	?	?	?	?	?	?	?	?	?	n	n	n	n	n	n	n
Craspedodiscus	n	n	n	0	n	n	3	3	?	1	?	0	n	n	n	n	n	n	n
Crucidenticula	n	0	n	0	n	n	0	n	n	n	n	0	0	2	2	0	0	3	1
Cussia	n	0	?	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Cyclotella	n	n	n	3	2	0	3	3	0	0	0	0	n	n	n	n	n	n	n
Cymatodiscus	n	0	n	?	?	?	?	?	?	?	?	?	n	n	n	n	n	n	n
Cymatogonia	n	0	n	?	?	?	?	?	?	?	?	?	n	n	n	n	n	n	n
Cymatopleura	n	0	n	0	n	n	0	n	n	n	n	0	1	2	2	0	0	3	0
Cymatosira	n	0	n	0	n	n	1	1	?	0	0	0	n	2	n	n	0	0	n
Cymatotheca	n	0	n	?	?	?	?	?	?	?	?	?	n	n	n	n	n	n	n
Cymbella	n	0	n	0	n	n	0	n	n	n	n	0	0	0	0	0	0	0	1
Dactyliosolen	n	n	n	0	n	n	1	1	?	0	0	n	n	n	n	n	n	n	n
Delphineis	n	0	n	0	n	n	2	6	2	0	0	0	0	1	0	v	0	0	n
Denticula	n	0	n	0	n	n	0	n	n	n	n	0	0	0	1	0	0	2	v
Denticulopsis	n	0	n	0	n	n	0	n	n	n	n	0	0	0	2	0	0	3	1
Dextradonator	n	1	0	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Diatoma	n	0	n	0	n	n	1	6	3	0	0	0	0	0	0	0	0	0	n

Table C.5: (continued)

Genus↓ *Character* →	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113
Dimerogramma	n	0	n	0	n	n	0	n	n	n	n	0	0	1	0	2	0	0	n
Diploneis	n	0	n	0	n	n	0	n	n	n	n	0	0	0	0	0	0	1	1
Discodiscus	n	n	n	0	n	n	3	1	?	?	?	0	n	n	n	n	n	n	n
Endictya	n	n	n	0	n	n	3	2	0	0	0	0	n	n	n	n	n	n	n
Epithemia	n	0	n	0	n	n	0	n	n	n	n	0	0	0	1	1	0	2	2
Ethmodiscus	n	n	n	0	n	n	3	v	3	v	0	0	n	n	n	n	n	n	n
Eucampia	n	0	0	0	n	n	1	1	?	0	0	0	n	n	n	n	n	n	n
Eunotia	n	0	n	0	n	n	1	6	0	0	0	0	0	0	2	0	0	4	1
Eunotogramma	n	0	n	0	n	n	1	1	?	?	?	0	n	n	n	n	n	n	n
Fenestrella	n	n	n	0	n	n	3	1	0	2	0	0	n	n	n	n	n	n	n
Fragilaria	n	0	n	0	n	n	1	6	0	0	0	0	0	0	0	0	0	0	n
Fragilariopsis	n	0	n	0	n	n	0	n	n	n	n	0	0	0	2	0	0	3	v
Gladiopsis	n	n	n	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Gladius	n	n	n	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Glyphodiscus	n	0	0	0	n	n	2	1	0	0	0	0	n	n	n	n	n	n	n
Gomphonema	0	0	n	0	n	n	0	n	n	n	n	0	0	0	0	0	0	1	1
Grammatophora	n	0	n	0	n	n	2	6	0	0	0	0	0	2	0	0	0	0	n
Grunowiella	n	0	n	0	n	n	1	6	?	?	?	0	0	0	0	2	0	0	n
Gyrosigma	n	0	n	0	n	n	0	n	n	n	n	0	0	0	0	0	0	1	1
Hemiaulus	n	0	0	0	n	n	1	0	?	?	?	0	n	n	n	n	n	n	n
Hemidiscus	n	0	n	0	n	n	3	2	0	1	2	0	n	n	n	n	n	n	n
Horodiscus	n	n	n	?	?	?	?	?	?	?	?	?	n	n	n	n	n	n	n
Huttonia	n	0	n	0	n	n	1	3	0	?	?	0	n	n	n	n	n	n	n
Hyalodiscus	n	n	n	0	n	n	3	1	0	0	0	0	n	n	n	n	n	n	n
Ikebea	?	0	n	?	?	?	?	?	?	?	?	?	0	0	0	0	0	0	n
Katathiraia	n	0	n	?	?	?	?	?	?	?	?	?	0	0	0	0	0	1	0
Kerkis	n	1	0	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n

Table C.5: (continued)

Genus↓ *Character* →	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113
Kisseleviella	n	0	n	0	n	n	1	2	0	0	0	0	n	n	n	n	n	n	n
Kozloviella	n	0	n	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Kreagra	n	n	n	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Liriogramma	n	n	n	0	n	n	3	2	0	2	2	0	n	n	n	n	n	n	n
Lisitzinia	n	0	n	?	?	?	?	?	?	?	?	?	n	n	n	n	n	n	n
Lithodesmium	n	0	0	0	n	n	1	0	1	0	0	0	n	n	n	n	n	n	n
Mastogloia	n	0	n	0	n	n	0	n	n	n	n	0	0	0	0	0	0	1	1
Mediaria	n	0	n	0	n	n	0	n	n	n	n	0	0	0	2	0	0	3	1
Melosira	n	n	n	0	n	n	3	1	2	0	0	0	n	n	n	n	n	n	n
Microorbis	n	n	n	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Monobrachia	n	0	n	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Navicula	n	0	n	0	n	n	0	n	n	n	n	0	0	0	0	0	1	1	1
Neobrunia	n	n	n	0	n	n	3	2	0	?	?	0	n	n	n	n	n	n	n
Neodelphineis	n	0	n	0	n	n	2	6	0	0	0	0	0	0	0	0	0	0	n
Neodenticula	n	0	n	0	n	n	0	n	n	n	n	0	0	0	2	0	0	3	0
Nitzschia	n	0	n	0	n	n	0	n	n	n	n	n	0	2	1	0	0	2	v
Odontella	n	0	0	0	n	n	2	0	1	0	?	0	n	n	n	n	n	n	n
Opephora	1	0	n	0	n	n	0	n	n	n	n	0	0	1	0	2	0	0	n
Paralia	n	n	n	0	n	n	3	3	0	0	0	0	n	n	n	n	n	n	n
Peponia	n	0	0	0	n	n	1	2	1	0	0	0	n	n	n	n	n	n	n
Plagiogramma	n	0	n	0	n	n	0	n	n	n	n	0	0	2	0	0	1	0	n
Planktoniella	n	n	n	3	3	0	1	3	0	?	?	0	n	n	n	n	n	n	n
Pleurosigma	n	0	n	0	n	n	0	n	n	n	n	0	0	0	0	0	0	1	1
Podosira	n	n	n	0	n	n	3	5	0	0	0	0	n	n	n	n	n	n	n
Porosira	n	n	n	3	1	0	1	2	2	0	0	0	n	n	n	n	n	n	n
Praethalassiosiropsis	n	n	n	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Pseudodimerogramma	n	0	n	0	n	n	1	6	?	?	?	0	0	0	0	0	0	0	n

Table C.5: (continued)

Genus↓ *Character* →	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113
Pseudoeunotia	?	0	?	?	?	?	?	?	?	?	?	?	0	2	?	?	0	0	n
Pseudopodosira	n	n	n	0	n	n	3	3	0	v	v	0	n	n	n	n	n	n	n
Pseudorutilaria	n	2	0	0	n	n	1	0	1	0	0	0	n	n	n	n	n	n	n
Pseudostictodiscus	n	n	n	?	?	?	?	?	?	?	?	?	n	n	n	n	n	n	n
Pseudotriceratium	n	v	0	0	n	n	3	6	0	0	0	0	n	n	n	n	n	n	n
Pyrgupyxis	n	n	n	?	?	?	?	?	?	?	?	?	n	n	n	n	n	n	n
Pyxilla	n	n	n	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Rattrayella	n	n	n	0	n	n	3	2	0	0	0	0	n	n	n	n	n	n	n
Rhabdonema	n	0	n	0	n	n	3	4	0	0	0	0	0	0	0	0	0	0	n
Rhaphidodiscus	n	0	n	0	n	n	0	n	n	n	n	0	0	0	0	0	0	1	1
Rhaphoneis	n	0	n	0	n	n	2	6	0	0	0	0	0	0	0	0	0	0	n
Rhizosolenia	n	n	n	0	n	n	1	0	3	?	?	0	n	n	n	n	n	n	n
Rhoicosphenia	0	0	n	0	n	n	0	n	n	n	n	0	0	0	0	2	0	1	1
Rhopalodia	n	0	n	0	n	n	0	n	n	n	n	0	0	0	0	2	0	3	1
Rhynchopyxis	n	n	n	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Riedelia	n	0	0	0	n	n	1	0	?	?	?	0	n	n	n	n	n	n	n
Rocella	n	n	n	0	n	n	1	0	0	0	0	0	n	n	n	n	n	n	n
Roperia	n	n	n	0	n	n	3	3	3	2	3	0	n	n	n	n	n	n	n
Rossiella	n	0	n	0	n	n	1	1	0	0	0	0	n	n	n	n	n	n	n
Rouxia	n	0	n	0	n	n	0	n	n	n	n	0	0	1	0	2	0	1	1
Rutilaria	n	0	n	0	n	n	1	0	1	0	0	0	n	n	n	n	n	n	n
Rylandsia	n	n	n	0	n	n	3	2	0	0	0	0	n	n	n	n	n	n	n
Sceptroneis	0	0	n	0	n	n	2	6	0	0	0	0	0	0	0	0	0	0	n
Simonseniella	n	n	n	0	n	n	?	?	?	?	?	0	n	n	n	n	n	n	n
Skeletonema	n	n	n	3	2	1	3	0	1	0	0	0	n	n	n	n	n	n	n
Sphynctolethus	n	1	0	0	n	n	1	0	1	0	0	0	n	n	n	n	n	n	n
Stellarima	n	n	n	0	n	n	3	0	0	0	0	0	n	n	n	n	n	n	n

Table C.5: (continued)

Genus↓ *Character* →	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113
Stephanodiscus	n	n	n	3	3	1	3	3	1	1	0	0	n	n	n	n	n	n	n
Stephanogonia	n	n	n	?	?	?	?	?	?	?	?	?	n	n	n	n	n	n	n
Stephanopyxis	n	n	n	0	n	n	3	2	1	?	?	0	n	n	n	n	n	n	n
Stictodiscus	n	0	n	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Strangulonema	n	n	n	?	?	?	?	?	?	?	?	?	n	n	n	n	n	n	n
Surirella	n	0	n	0	n	n	0	n	n	n	n	0	0	1	0	2	0	3	1
Synedra	n	0	n	0	n	n	2	6	0	0	0	0	0	0	0	0	1	0	n
Tetracyclus	n	0	n	0	n	n	v	0	0	0	0	0	0	0	0	0	0	0	n
Thalassionema	n	2	n	0	n	n	2	6	0	0	0	0	0	1	0	2	0	0	n
Thalassiosira	n	n	n	3	3	0	1	3	1	0	0	0	n	n	n	n	n	n	n
Thalassiothrix	n	3	n	0	n	n	2	6	0	0	3	0	0	1	0	2	0	0	n
Trachyneis	n	0	n	0	n	n	0	n	n	n	n	0	0	2	0	0	0	1	1
Triceratium	n	0	1	0	n	n	3	2	1	?	?	0	n	n	n	n	n	n	n
Trinacria	n	2	0	0	n	n	3	0	1	0	0	0	n	n	n	n	n	n	n
Trochosira	n	n	n	0	n	n	3	2	0	0	0	0	n	n	n	n	n	n	n
Trochus	n	n	n	0	n	n	0	n	n	n	n	0	n	n	n	n	n	n	n
Tropidoneis	n	0	n	0	n	n	0	n	n	n	n	0	0	0	1	0	0	2	1

Table C.6: Morphospace data matrix (characters 114-123)

Genus↓ *Character* →	114	115	116	117	118	119	120	121	122	123
Abas	n	n	n	n	n	n	n	n	n	n
Achnanthes	0	0	0	?	1	v	0	0	0	0
Actinocyclus	n	n	0	n	n	n	n	n	n	n
Actinoptychus	n	n	0	n	n	n	n	n	n	n
Amblypyrgus	n	n	n	n	n	n	n	n	n	n
Amphora	0	2	1	3	v	0	0	0	0	0

Table C.6: (continued)

Genus↓ *Character* →	114	115	116	117	118	119	120	121	122	123
Anaulus	n	n	n	n	n	n	n	n	n	n
Ancylopyrgus	n	n	n	n	n	n	n	n	n	n
Annellus	n	n	o	n	n	n	n	n	n	n
Arachnoidiscus	n	n	n	n	n	n	n	n	n	n
Archepyrgus	n	n	n	n	n	n	n	n	n	n
Asterolampra	n	n	n	n	n	n	n	n	n	n
Asteromphalus	n	n	n	n	n	n	n	n	n	n
Aulacodiscus	n	n	n	n	n	n	n	n	n	n
Aulacoseira	n	n	n	n	n	n	n	n	n	n
Azpeitia	n	n	n	n	n	n	n	n	n	n
Bacillaria	o	o	o	v	n	n	o	1	1	o
Bacteriastrum	n	n	n	n	n	n	n	n	n	n
Bacterosira	n	n	n	n	n	n	n	n	n	n
Basilicostephanus	n	n	n	n	n	n	n	n	n	n
Baxteriopsis	n	n	n	n	n	n	n	n	n	n
Biddulphia	n	n	n	n	n	n	n	n	n	n
Bilingua	n	n	n	n	n	n	n	n	n	n
Bogorovia	n	n	n	n	n	n	n	n	n	n
Brightwellia	n	n	n	n	n	n	n	n	n	n
Caloneis	o	1	o	6	1	o	o	o	o	o
Campyloneis	o	o	o	o	o	o	o	o	o	o
Ceratoneis	o	o	o	o	1	o	1	1	1	?
Cereataulus	n	n	n	n	n	n	n	n	n	n
Cestodiscus	n	n	n	n	n	n	n	n	n	n
Chaetoceros	n	n	n	n	n	n	n	n	n	n
Charcotia	n	n	n	n	n	n	n	n	n	n
Cladogramma	n	n	n	n	n	n	n	n	n	n

Table C.6: (continued)

Genus↓ *Character* →	114	115	116	117	118	119	120	121	122	123
Clavicula	n	n	n	n	n	n	n	n	n	n
Cocconeis	0	0	0	0	1	0	0	1	0	0
Corethron	n	n	n	n	n	n	n	n	n	n
Coscinodiscus	n	n	n	n	n	n	n	n	n	n
Cosmiodiscus	n	n	n	n	n	n	n	n	n	n
Craspedodiscus	n	n	n	n	n	n	n	n	n	n
Crucidenticula	0	0	0	?	?	?	1	0	0	1
Cussia	n	n	n	n	n	n	n	n	n	n
Cyclotella	n	n	n	n	n	n	n	n	n	n
Cymatodiscus	n	n	n	n	n	n	n	n	n	n
Cymatogonia	n	n	n	n	n	n	n	n	n	n
Cymatopleura	1	0	0	0	n	n	0	1	1	0
Cymatosira	n	n	0	n	n	n	n	n	n	n
Cymatotheca	n	n	n	n	n	n	n	n	n	n
Cymbella	0	v	0	5	1	0	0	0	0	0
Dactyliosolen	n	n	n	n	n	n	n	n	n	n
Delphineis	n	n	n	n	n	n	n	n	n	n
Denticula	0	0	0	6	v	?	1	1	1	0
Denticulopsis	0	0	0	0	1	0	1	0	0	0
Dextradonator	n	n	n	n	n	n	n	n	n	n
Diatoma	n	n	n	n	n	n	n	n	n	n
Dimerogramma	n	n	n	n	n	n	n	n	n	n
Diploneis	0	0	0	6	v	0	0	0	0	0
Discodiscus	n	n	n	n	n	n	n	n	n	n
Endictya	n	n	n	n	n	n	n	n	n	n
Epithemia	0	2	0	0	1	0	1	0	1	0
Ethmodiscus	n	n	n	n	n	n	n	n	n	n

Table C.6: (continued)

Genus↓ *Character* →	114	115	116	117	118	119	120	121	122	123
Eucampia	n	n	n	n	n	n	n	n	n	n
Eunotia	0	2	0	6	1	0	0	0	0	0
Eunotogramma	n	n	n	n	n	n	n	n	n	n
Fenestrella	n	n	n	n	n	n	n	n	n	n
Fragilaria	n	n	n	n	n	n	n	n	n	n
Fragilariopsis	0	0	0	0	0	0	1	0	1	0
Gladiopsis	n	n	n	n	n	n	n	n	n	n
Gladius	n	n	n	n	n	n	n	n	n	n
Glyphodiscus	n	n	n	n	n	n	n	n	n	n
Gomphonema	0	2	0	4	0	0	0	0	0	0
Grammatophora	n	n	n	n	n	n	n	n	n	n
Grunowiella	n	n	n	n	n	n	n	n	n	n
Gyrosigma	0	1	0	5	5	0	0	0	0	0
Hemiaulus	n	n	n	n	n	n	n	n	n	n
Hemidiscus	n	n	n	n	n	n	n	n	n	n
Horodiscus	n	n	n	n	n	n	n	n	n	n
Huttonia	n	n	n	n	n	n	n	n	n	n
Hyalodiscus	n	n	n	n	n	n	n	n	n	n
Ikebea	n	n	n	n	n	n	n	n	n	n
Katathiraia	0	0	0	?	n	n	1	0	?	0
Kerkis	n	n	n	n	n	n	n	n	n	n
Kisseleviella	n	n	n	n	n	n	n	n	n	n
Kozloviella	n	n	n	n	n	n	n	n	n	n
Kreagra	n	n	n	n	n	n	n	n	n	n
Liriogramma	n	n	n	n	n	n	n	n	n	n
Lisitzinia	n	n	n	n	n	n	n	n	n	n
Lithodesmium	n	n	n	n	n	n	n	n	n	n

Table C.6: (continued)

Genus↓ *Character* →	114	115	116	117	118	119	120	121	122	123
Mastogloia	0	2	1	6	0	0	0	0	0	0
Mediaria	0	0	0	0	0	0	0	0	0	0
Melosira	n	n	n	n	n	n	n	n	n	n
Microorbis	n	n	n	n	n	n	n	n	n	n
Monobrachia	n	n	n	n	n	n	n	n	n	n
Navicula	0	0	1	v	v	0	0	0	0	1
Neobrunia	n	n	n	n	n	n	n	n	n	n
Neodelphineis	n	n	n	n	n	n	n	n	n	n
Neodenticula	0	0	0	?	n	n	1	0	0	0
Nitzschia	0	0	1	v	4	0	1	1	1	0
Odontella	n	n	n	n	n	n	n	n	n	n
Opephora	n	n	n	n	n	n	n	n	n	n
Paralia	n	n	n	n	n	n	n	n	n	n
Peponia	n	n	n	n	n	n	n	n	n	n
Plagiogramma	n	n	n	n	n	n	n	n	n	n
Planktoniella	n	n	n	n	n	n	n	n	n	n
Pleurosigma	0	1	0	6	4	0	0	0	0	0
Podosira	n	n	n	n	n	n	n	n	n	n
Porosira	n	n	n	n	n	n	n	n	n	n
Praethalassiosiropsis	n	n	n	n	n	n	n	n	n	n
Pseudodimerogramma	n	n	n	n	n	n	n	n	n	n
Pseudoeunotia	n	n	n	n	n	n	n	n	n	n
Pseudopodosira	n	n	n	n	n	n	n	n	n	n
Pseudorutilaria	n	n	n	n	n	n	n	n	n	n
Pseudostictodiscus	n	n	n	n	n	n	n	n	n	n
Pseudotriceratium	n	n	n	n	n	n	n	n	n	n
Pyrgupyxis	n	n	n	n	n	n	n	n	n	n

Table C.6: (continued)

Genus↓ *Character* →	114	115	116	117	118	119	120	121	122	123
Pyxilla	n	n	n	n	n	n	n	n	n	n
Rattrayella	n	n	n	n	n	n	n	n	n	n
Rhabdonema	n	n	n	n	n	n	n	n	n	n
Rhaphidodiscus	0	0	0	0	0	0	0	0	0	0
Rhaphoneis	n	n	n	n	n	n	n	n	n	n
Rhizosolenia	n	n	n	n	n	n	n	n	n	n
Rhoicosphenia	0	0	0	4	1	0	0	0	0	0
Rhopalodia	0	0	0	0	1	0	1	1	1	0
Rhynchopyxis	n	n	n	n	n	n	n	n	n	n
Riedelia	n	n	n	n	n	n	n	n	n	n
Rocella	n	n	n	n	n	n	n	n	n	n
Roperia	n	n	n	n	n	n	n	n	n	n
Rossiella	n	n	n	n	n	n	n	n	n	n
Rouxia	0	0	0	0	0	0	0	0	0	0
Rutilaria	n	n	n	n	n	n	n	n	n	n
Rylandsia	n	n	n	n	n	n	n	n	n	n
Sceptroneis	n	n	n	n	n	n	n	n	n	n
Simonseniella	n	n	n	n	n	n	n	n	n	n
Skeletonema	n	n	n	n	n	n	n	n	n	n
Sphynctolethus	n	n	n	n	n	n	n	n	n	n
Stellarima	n	n	n	n	n	n	n	n	n	n
Stephanodiscus	n	n	n	n	n	n	n	n	n	n
Stephanogonia	n	n	n	n	n	n	n	n	n	n
Stephanopyxis	n	n	n	n	n	n	n	n	n	n
Stictodiscus	n	n	n	n	n	n	n	n	n	n
Strangulonema	n	n	n	n	n	n	n	n	n	n
Surirella	1	0	0	0	n	n	1	1	1	0

Table C.6: (continued)

Genus↓ *Character* →	114	115	116	117	118	119	120	121	122	123
Synedra	n	n	n	n	n	n	n	n	n	n
Tetracyclus	n	n	n	n	n	n	n	n	n	n
Thalassionema	n	n	n	n	n	n	n	n	n	n
Thalassiosira	n	n	n	n	n	n	n	n	n	n
Thalassiothrix	n	n	n	n	n	n	n	n	n	n
Trachyneis	0	0	0	6	4	1	0	0	0	1
Triceratium	n	n	n	n	n	n	n	n	n	n
Trinacria	n	n	n	n	n	n	n	n	n	n
Trochosira	n	n	n	n	n	n	n	n	n	n
Trochus	n	n	n	n	n	n	n	n	n	n
Tropidoneis	0	2	0	v	?	0	0	1	0	0

D

Sources of Morphological Descriptions

Table D.1: Sources of morphological descriptions for diatom morphospace

Genus	Source
Abas	Round et al. (1990)
Achnanthes	Round et al. (1990); Cox (2006)
Actinocyclus	Round et al. (1990)
Actinoptychus	Round et al. (1990)
Amblypyrgus	Gersonde and Harwood (1990)
Amphora	Round et al. (1990)
Anaulus	Round et al. (1990); Drebes and Schulz (1989)
Ancylopyrgus	Gersonde and Harwood (1990)
Annellus	Burckle (1974); Barron (1976, 1981)
Arachnoidiscus	Round et al. (1990)
Archepyrgus	Gersonde and Harwood (1990)
Asterolampra	Round et al. (1990)
Asteromphalus	Round et al. (1990)
Aulacodiscus	Round et al. (1990)
Aulacoseira	Round et al. (1990)
Azpeitia	Round et al. (1990)
Bacillaria	Round et al. (1990); Schmid (2007)

Table D.1: (continued)

Genus	Source
Bacteriastrum	Round et al. (1990)
Bacterosira	Guiry and Guiry (2011); Alverson (2007)
Basilicostephanus	Gersonde and Harwood (1990)
Baxteriopsis	Van Heurck and Baxter (1896); Bolli et al. (1989); Barron and Mahood (1993)
Biddulphia	Round et al. (1990)
Bilingua	Gersonde and Harwood (1990)
Bogorovia	Yanagisawa (1995a)
Brightwellia	Round et al. (1990)
Caloneis	Round et al. (1990)
Campyloneis	Round et al. (1990); De Stefano et al. (2003)
Ceratoneis	Jahn and Kusber (2005)
Cereataulus	Akiba (1986)
Cestodiscus	Greville (1863); Fenner (1984); Tuji et al. (2009)
Chaetoceros	Round et al. (1990)
Charcotia	Simonsen (1982)
Cladogramma	Van Heurck and Baxter (1896); Moshkovitz et al. (1983); Suto et al. (2009)
Clavicula	Van Heurck and Baxter (1896)
Cocconeis	Round et al. (1990)
Corethron	Round et al. (1990); Llano and Wallen (1971)
Coscinodiscus	Round et al. (1990)
Cosmiodiscus	Koizumi (1973); Bolli et al. (1989, p. 783)
Craspedodiscus	Round et al. (1990)
Crucidenticula	Akiba (1986)
Cussia	Schrader (1974)
Cyclotella	Round et al. (1990)
Cymatodiscus	Hendey (1958)
Cymatogonia	Hanna (1932)
Cymatopleura	Round et al. (1990); Spaulding and Edlund (2009)
Cymatosira	Round et al. (1990)
Cymatotheca	Hendey (1958)
Cymbella	Round et al. (1990)
Dactyliosolen	Round et al. (1990); Hasle (1975)
Delphineis	Round et al. (1990)

Table D.1: (continued)

Genus	Source
Denticula	Round et al. (1990)
Denticulopsis	Yanagisawa and Akiba (1990); Simonsen (1979)
Dextradonator	Ross and Sims (1980)
Diatoma	Round et al. (1990)
Dimerogramma	Round et al. (1990)
Diploneis	Round et al. (1990)
Discodiscus	Gombos (1980)
Endictya	Round et al. (1990)
Epithemia	Spaulding and Edlund (2010)
Ethmodiscus	Round et al. (1990)
Eucampia	Round et al. (1990)
Eunotia	Round et al. (1990)
Eunotogramma	Round et al. (1990)
Fenestrella	Sims (1990)
Fragilaria	Round et al. (1990)
Fragilariopsis	Round et al. (1990)
Gladiopsis	Gersonde and Harwood (1990)
Gladius	Gersonde and Harwood (1990)
Glyphodiscus	Stidolph (1985)
Gomphonema	Round et al. (1990)
Grammatophora	Round et al. (1990)
Grunowiella	Van Heurck and Baxter (1896); Fenner (1991a); Sims et al. (2006)
Gyrosigma	Round et al. (1990)
Hemiaulus	Round et al. (1990)
Hemidiscus	Round et al. (1990)
Horodiscus	Hanna (1927)
Huttonia	Garcia (2004)
Hyalodiscus	Round et al. (1990)
Ikebea	Komura (1975); Olney et al. (2007)
Katathiraia	Komura (1976)
Kerkis	Gersonde and Harwood (1990)
Kisseleviella	Olney et al. (2005); Sheshukova-Poretskaya (1962)
Kozloviella	Jousé (1974); Fenner (1984)
Kreagra	Gersonde and Harwood (1990)

Table D.1: (continued)

Genus	Source
Liriogramma	Round et al. (1990)
Lisitzinia	Gombos Jr and Ciesielski (1983); Bolli et al. (1989, pp. 734 & 748)
Lithodesmium	Round et al. (1990)
Mastogloia	Round et al. (1990)
Mediaria	Yanagisawa (1994); Bolli et al. (1989, p. 787)
Melosira	Round et al. (1990)
Microorbis	Gersonde and Harwood (1990)
Monobrachia	Schrader and Fenner (1976)
Navicula	Round et al. (1990)
Neobrunia	Hendey (1981); Van Heurck and Baxter (1896)
Neodelphineis	Round et al. (1990)
Neodenticula	Akiba (1986)
Nitzschia	Round et al. (1990)
Odontella	Round et al. (1990)
Opephora	Sabbe and Vyverman (1995)
Paralia	Round et al. (1990)
Peponia	Olshtynskaya (2002)
Plagiogramma	Round et al. (1990)
Planktoniella	Round et al. (1990)
Pleurosigma	Round et al. (1990)
Podosira	Round et al. (1990)
Porosira	Round et al. (1990)
Praethalassiosiropsis	Gersonde and Harwood (1990)
Pseudodimerogramma	Schrader and Fenner (1976)
Pseudoeunotia	Hustedt and Jensen (1985); Schrader (1974); Bolli et al. (1989, p. 788)
Pseudopodosira	Olshtynskaja and Simola (1990)
Pseudorutilaria	Round et al. (1990); Ross and Sims (1987)
Pseudostictodiscus	Schrader and Fenner (1976); Scherer and Koç (1996); Baldauf and Barron (1987)
Pseudotriceratium	Round et al. (1990)
Pyrgupyxis	Hendey (1969)
Pyxilla	Round et al. (1990)
Rattrayella	Sims (2006)

Table D.1: (continued)

Genus	Source
Rhabdonema	Round et al. (1990)
Rhaphidodiscus	Andrews (1988); Bolli et al. (1989, p. 736)
Rhaphoneis	Round et al. (1990)
Rhizosolenia	Round et al. (1990)
Rhoicosphenia	Round et al. (1990)
Rhopalodia	Round et al. (1990)
Rhynchopyxis	Gersonde and Harwood (1990)
Riedelia	Schrader and Fenner (1976, p. 997)
Rocella	Round et al. (1990)
Roperia	Round et al. (1990)
Rossiella	Yanagisawa (1995b)
Rouxia	Akiba (1986); Hanna (1930)
Rutilaria	Round et al. (1990)
Rylandsia	Gombos (1980)
Sceptroneis	Round et al. (1990)
Simonseniella	Fenner (1991b); Akiba (1986)
Skeletonema	Round et al. (1990)
Sphynctolethus	Sims (1986)
Stellarima	Hasle and Sims (1986)
Stephanodiscus	Round et al. (1990)
Stephanogonia	Andrews (1976); Hanna (1932); Hajós (1976)
Stephanopyxis	Round et al. (1990)
Stictodiscus	Round et al. (1990)
Strangulonema	Round et al. (1990)
Surirella	Round et al. (1990)
Synedra	Round et al. (1990)
Tetracyclus	Round et al. (1990)
Thalassionema	Round et al. (1990)
Thalassiosira	Round et al. (1990)
Thalassiothrix	Round et al. (1990)
Trachyneis	Round et al. (1990)
Triceratium	Round et al. (1990)
Trinacria	Round et al. (1990)
Trochosira	Sims (1988)
Trochus	Gersonde and Harwood (1990)

Table D.1: (continued)

Genus	Source
Tropidoneis	Patrick and Reimer (1975)

E

Morphospace Characters Grouped by Evolutionary Hypothesis

Table E.1: Characters expected to relate to changes in predation

Suggesting more predation resistance	Suggesting less predation resistance
Character 1, state 0	States 1, 2, 3, 4
Character 2, state 4	States 0, 1, 2, 3, 6, 7
Character 12, state 2	States 0, 1, 3, 4, 5
Character 27, state 2	States 0, 1
Character 37, state 1	State 0
Character 43, states 1, 2, 3	States 0, 4, 5
Character 86, states 1, 2	State 0
Character 87, state 1, 2	State 0
Character 88, state 1	State 0
Character 89, states 1, 2	State 0
Character 90, state 1	State 0
Character 121, state 1	State 0
Character 122, state 1	State 0

Table E.2: Characters expected to relate to cell-cell linkage

Suggesting cell-cell linkage present	Suggesting cell-cell linkage absent
Character 18, states 1, 2, 3	State 0
Character 27, states 1, 2	State 0
Character 28, state 1	State 0
Character 55, states 1, 2	State 0
Character 59, states 1, 2	State 0
Character 96, states 1, 2	State 0
Character 98, state 3	State 0

Table E.3: Characters expected to relate to viral defense

Sugg. more protection from viral attack	Sugg. less protection from viral attack
Character 41, states 1, 2, 3, 4	State 0
Character 68, state 0	States 1, 2, 3
Character 69, state 5	States 0, 1, 2, 3, 4
Character 70, state 2	States 0, 1

Table E.4: Characters expected to relate to changes in silica availability

Suggesting more silica use	Suggesting less silica use
Character 61, state 0	States 1, 2, 3
Character 59, state 0	States 1, 2
Character 70, state 0	States 1, 2
Character 80, state 1	State 0
Character 98, state 3	State 0
Character 27, state 2	States 0, 1
Character 29, state 0	States 1, 2, 3
Character 35, state 0	State 1
Character 36, state 0	State 1
Character 37, state 0	State 1
Character 38, states 0, 2, 3	State 5
Character 43, state 0	States 1, 2, 3, 4, 5
Character 49, state 0	State 1
Character 53, states 0, 1	States 2, 4
Character 55, state 0	States 1, 2
Character 56, state 0	State 1
Character 83, state 0	State 1
Character 84, state 0	State 1

F

R Code for Morphospace Analysis

F.1 R script for analysis and plotting

Code F.1: Morphospace.R

```
#Ben Kotrc, Harvard University, November 2011
#kotrc@fas.harvard.edu
#R routine to carry out analysis of diatom morphospace
#Makes calls to functions stored in "MorphospaceFunctions.R"
#and in "DiversityFunctions.R"

#Load in packages, data, and functions needed
library("scatterplot3d");
library("grid");
library("ReadImages");
library("ape");
library("shape");
library("MASS");
library("cluster");
library("vcd");
library("RColorBrewer");
library("colorspace");
library("seqinr");
library("YaleToolkit");
library("grImport");
library("png");
library("geometry");
library("alphashape3d");

#Read in the raw data matrix
```

Code F.1: Morphospace.R (continued)

```
mfull <- read.table(file='Matrix.txt', header=TRUE, sep=",", row.names=1,colClasses="
    character");
#Read in culled data matrix (to 80% completeness threshold)
m <- read.table(file="Matrix80Cull.txt", sep=",",header=TRUE);
#Read distance matrix based on data culled to >80% using "?" only as missing
d <- as.matrix(read.table(file='DistanceMatrixCulled8080?nosingles.txt', header=TRUE, row.
    names=1, sep=",",as.is=TRUE));
#Read in Neptune database, with Cretaceous occurrences added, and Genus name field,
#and with typos/mistakes/synonyms corrected
N <- read.table(file='NeptuneGenNamesCorrCret.txt', header=TRUE, sep="\t",as.is=TRUE);
#For paper 2 (subsampling), throw out the added Cretaceous stuff---Neptune only
N <- N[N$Sample.Age < 65,];
#Source file containing the morphospace analysis functions
source("MorphospaceFunctions.R");

#A note on embedding fonts in PDF output figures:
#The publication figures use Gill Sans. For the related code to work
#.OTF and .AFM font files need to be placed in the appropriate directories
#and Ghostscript (gs) needs to be installed at the Terminal command line. That in
#turn is best done with Fink. Even when gs is installed, the R function that
#calls gs through the system may struggle to find it.
#If that happens, make a symbolic link from /usr/bin pointing to gs directly;
#something like: ln -s /sw/bin/gs ./gs
#(use commmand 'which gs' to find where gs lives)

###DATA QUALITY R^2 PLOT (PUBLICATION 2)###
#Sensitivity of results to thresholds of data culling
#Using only "?" as a definition of missing data
#Assess the completeness (genera and characters) of the matrix
dq <- getDataQuality(mfull, inv="?");
#Characters range from 67%-100% complete, genera from 57%
#Go through each level that's in the data
complevs <- sort(unique(c(floor(unique(dq$gen)),floor(unique(dq$char)))))
#For each completeness level, we're going to want an r^2 value
#for each disparity metric, ready vector to hold them
mpwd <- vector("numeric",length=length(complevs));
cvol <- mpwd;
rsq <- as.data.frame(cbind(mpwd,cvol));
rownames(rsq) <- complevs;
pvals <- rsq;
#Which taxon sampling mode to use?
#Options: "in-bin", "range-through","uw","cr","sqs"
samplingmode <- "in-bin";
#(N.b.: because 65 myrs is an odd number, the youngest interval is only 1 myr long)
bins <- c(seq(from=62,to=2,by=-2));
#Set time bin names
binnames <- c(seq(from=64,to=2,by=-2)-1);
#Read distance matrix based on data culled to >80% using "?" only as missing
#(as used in paper 1)
dref <- as.matrix(read.table(file='DistanceMatrixCulled8080?nosingles.txt', header=TRUE,
    row.names=1, sep=",",as.is=TRUE))
#We're going to need a reference set of disparity results
#to compare to---using the 80% culling threshold from paper 1
dispref <- plotDivDispPub2LgDQ(bins, binnames, samplingmode, N, dref, sptrials=1,gentrials
    =1,sendback=TRUE)
#Loop through each completeness level
for (i in 1:length(complevs))
{
```

Code F.1: Morphospace.R (continued)

```
  cat("Currently on ",complevs[i],"%, ",length(complevs)-i," to go\n")
  #Make a copy of the full morphospace matrix for culling
  mcull <- mfull;
  #Keep track of the size of the matrix, to stop when it's stable
  msize <- c(1000,1000);
  while(!identical(dim(mcull),msize)){
    #Update matrix size
    msize <- dim(mcull)
    #Cull by ith completeness level percentage
    #mcull <- cullMatrix(mcull,getDataQuality(mcull, inv="?"),complevs[i],complevs[i]);
    #Characters=100
    mcull <- cullMatrix(mcull,getDataQuality(mcull, inv="?"),0,complevs[i]);
    #Genus=100
    mcull <- cullMatrix(mcull,getDataQuality(mcull, inv="?"),complevs[i],0);
    #Now throw out "uninformative" characters with less than two
    #states having one or more valid entries
    uninf <- vector(length=ncol(mcull));
    names(uninf) <- colnames(mcull);
    for(j in 1:ncol(mcull))
    {
      #If the character has less than 2 states with more than 1 valid entry
      if(sum(table(as.factor(makeNumeric(mcull[,j]))) > 1) < 2)
      {
        #Flag as uninformative
        uninf[j] <- TRUE;
      }
    }
    #Cull the matrix to take out those "uninformative" characters
    mcull <- mcull[,!uninf];
  }
  #Now that we have a culled matrix, let's run the disparity metrics
  #First, get distance matrix
  dcull <- getDistMatrix(mcull);
  #Throw out Neptune occurrences that aren't in the matrix
  #(this step is crucial or it will break)
  Ncull <- N[N$Genus %in% row.names(dcull),];
  #Now run in-bin disparity metrics using this matrix
  dispcull <- plotDivDispPub2LgDQ(bins, binnames, samplingmode, Ncull, dcull, sptrials
       =1000,gentrials=1000,sendback=TRUE);
  #Get correlation and p-value for reference (x) against ith iteration data
  #for each of the metrics
  rsq$mpwd[i] <- cor(dispref$mpwd,dispcull$mpwd,use="complete.obs");
  rsq$cvol[i] <- cor(dispref$chullvols,dispcull$chullvols,use="complete.obs");
  pvals$mpwd[i] <- cor.test(dispref$mpwd,dispcull$mpwd,use="complete.obs")$p.value;
  pvals$cvol[i] <- cor.test(dispref$chullvols,dispcull$chullvols,use="complete.obs")$p.
       value;
  #Save to file so we can see where it crashes
  write.table(rsq[1:i,],file="dataquality_r-squareds.txt");
  write.table(pvals[1:i,],file="dataquality_p-vals.txt");
}
#Plotting:
#Get the fonts ready
file.exists <- function( fname ) length(Sys.glob(fname))>0
absolute.path.to.font.files <-  "/Users/bkotrc/font/";
bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
## if you do not have the correct font types
for (i in 1:length(bera.names)) {
  stopifnot( file.exists(paste(absolute.path.to.font.files,
```

Code F.1: Morphospace.R (continued)

```
                              bera.names[i], ".afm", sep="")) )
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                              bera.names[i], ".otf", sep="")) )
}
gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files,bera.names, ".afm", sep
    =""))
pdfname <- "dataculling.pdf";
#Make a composite plot of previous plus this plot for publication
pdf(file=pdfname, bg="white", width=(7.2/cm(1)), height=(14.4/cm(1)), pointsize=7, family=
    gillsans);
par(font.main=1, cex.main=1.5)
axthck <- 0.3;
par(mfrow=c(2,1), lwd=axthck);
plot(row.names(rsq),rsq$cvol^2,bty="n",type="n",axes=FALSE,xlab="Data quality threshold",
    ylab="Correlation with reference results",main="",lwd=0.5,xaxs="i",yaxs="i",ylim=c
    (0.6,1),xlim=c(50,100));
mtext("(% of character states not coded as '?', for characters and genera)",cex=0.75,side
    =1,line=4,font=3)
mtext(expression((R^2)),cex=0.75,side=2,line=2,font=3)
title(main="Convex hull volume",adj=1);
text(x=79.5,y=0.65,labels=paste("Threshold used to \n obtain reference results"),adj=1,cex
    =.75,font=3)
points(row.names(rsq),rsq$cvol^2,pch=3,lwd=0.8,xpd=TRUE);
abline(v=80,lty=2)
points(x=c(50,80),y=c(1,1),type="l",lty=2,lwd=axthck)
axis(1, lwd.ticks=axthck, lwd=axthck);
axis(2, lwd=axthck);
plot(row.names(rsq),rsq$mpwd^2,bty="n",type="n",axes=FALSE,xlab="Data quality threshold",
    ylab="Correlation with reference results",main="",lwd=0.5,xaxs="i",yaxs="i",ylim=c
    (0.5,1),xlim=c(50,100));
title(main="Mean pairwise distance",adj=1);
points(row.names(rsq),rsq$mpwd^2,pch=3,lwd=0.8,xpd=TRUE);
abline(v=80,lty=2)
points(x=c(50,80),y=c(1,1),type="l",lty=2,lwd=axthck)
axis(1, lwd.ticks=axthck, lwd=axthck);
axis(2, lwd=axthck);
#Now embed font in that file
dev.off()
embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font");
######


###DATA QUALITY DETAILED COMPARISON PLOT (PUBLICATION 2)###
#How different are the results under 80% vs 100% completeness
#thresholds? The R^2 is about 0.6, but what does that look like?
#Source file containing the morphospace analysis functions
source("MorphospaceFunctions.R")
#Do the thing
plotDQ80100Comparison()
######


###CULL MATRIX (FOR PUBLICATION 1)###
#Subset the existing matrix, mfull, by selecting only genera
#and characters meeting a threshold %age of completeness as calculated above
#mcull <- cullMatrix(mfull,getDataQuality(mfull),50,50);
mcull <- mfull;
#First, throw out characters less than 20% complete (arbitrary choice)
```

Code F.1: Morphospace.R (continued)

```
mcull <- cullMatrix(mcull,getDataQuality(mcull, inv="?"),0,80);
#Next, throw out genera less than 50% complete (also arbitrary)
mcull <- cullMatrix(mcull,getDataQuality(mcull, inv="?"),80,0);
#Now, see if there are any 'uninformative' characters left
#(again, arbitrarily, characters not having at least two states with more than
#one valid entry each)
uninf <- vector(length=ncol(mcull));
names(uninf) <- colnames(mcull);
for(i in 1:ncol(mcull))
{
  #If the character has less than 2 states with more than 1 valid entry
  if(sum(table(as.factor(makeNumeric(mcull[,i]))) > 1) < 2)
  {
    #Flag as uninformative
    uninf[i] <- TRUE;
  }
}
#Cull the matrix to take out those "uninformative" characters
mcull <- mcull[,!uninf];
#Repeat this until the matrix dimensions are stable (140 x 100), i.e. 2x
#Remove character X76, pseudoloculate, which arguably doesn't represent
#a real morphological difference (rather than a developmental one)
#Since this is a bit of a hassle, save to disk
write.table(as.matrix(mcull),file="Matrix80Cull.txt", sep=",",col.names=TRUE, row.names=
    TRUE);
#How sparse is the full data matrix?
sparse <- sum(mfull == 'n' | mfull == 'v' | mfull == '?')/(dim(mfull)[1]*dim(mfull)[2])
#And the culled?
sparse <- sum(mcull == 'n' | mcull == 'v' | mcull == '?')/(dim(mcull)[1]*dim(mcull)[2])
#Compare to Foote's matrices
#Read his file (modified so it's just the matrix)
crinoids <- read.fwf(file="Foote_Crinoid_Modified.txt",widths=rep(1,90));
sparse <- sum(crinoids == 'n' | crinoids == 'v' | crinoids == '?')/(dim(crinoids)[1]*dim(
    crinoids)[2])
#Again
crinoids <- read.fwf(file="Foote_Blasto_Modified.txt",widths=rep(1,65));
sparse <- sum(crinoids == 'n' | crinoids == 'v' | crinoids == '?')/(dim(crinoids)[1]*dim(
    crinoids)[2])
######


###CALCULATE DISTANCE/DISSIMILARITY MATRIX, D###
#Which morphospace matrix to use for calculation?
m <- mcull;
#Calculate pairwise distances between genera based on ratio of
#character state matches to possible matches
d <- getDistMatrix(m);
#Save dissimilarity matrix to file
write.table(as.matrix(d),file="DistanceMatrixCulled8080?nosingles.txt", sep=",",col.names=
    TRUE, row.names=TRUE);
#Load saved dissimilarity matrix from file
#Either: matrix culled to >50% using "?,n,v" as missing
#d <- as.matrix(read.table(file='DistanceMatrixCulled50%.txt', header=TRUE, row.names=1,
    sep=",",as.is=TRUE));
#Or: matrix culled to >80% using "?" only as missing
d <- as.matrix(read.table(file='DistanceMatrixCulled8080?nosingles.txt', header=TRUE, row.
    names=1, sep=",",as.is=TRUE));
######
```

Code F.1: Morphospace.R (continued)

```
238
239
240 ###PCO FIT/VARIANCE EXPLAINED, BY EIGENVALUES###
241 #Run PCO with chosen number of k eigenvectors retained
242 pco <- cmdscale(d, k=2, eig=TRUE);
243 #Interrogate GOF ("goodness of fit")
244 pco$GOF;
245 #All eigenvalues sum to:
246 sum(pco$eig);
247 #Of this, the first and second eigenvalues are 18% and 12%
248 pco$eig[1]/sum(pco$eig);
249 pco$eig[2]/sum(pco$eig);
250 #Positive eigenvalues sum to
251 sum(pco$eig[pco$eig>0]);
252 #Of this, the first and second eigenvalues are 12% and 9%
253 pco$eig[1]/sum(pco$eig[pco$eig>0]);
254 pco$eig[2]/sum(pco$eig[pco$eig>0]);
255 #(To save as PDF, uncomment next line)
256 pdf(file='EigenvaluesCulled80?.pdf', width=6, height=6);
257 #Let's take a look at the eigenvalues plotted, that should give a sense
258 plot(1:length(pco$eig),pco$eig,bty="n",xlab="Eigenvalue number",ylab="Eigenvalue",main="
        Eigenvalues returned by cmdscale()");
259 mtext('Using data matrix culled for genera and characters with < 20% "?" states', side=3,
        cex=0.8);
260 abline(h=0);
261 dev.off();
262 #Compare GOF eigenvalues with those from correlation above:
263 gofs1 <- c(0,0,0,0);
264 gofs2 <- c(0,0,0,0);
265 eigfracs <- c(0,0,0,0);
266 for(i in 1:4)
267 {
268   pco <- cmdscale(d, k=i, eig=TRUE);
269   gofs1[i] <- pco$GOF[1];
270   gofs2[i] <- pco$GOF[2];
271   eigfracs[i] <- sum(pco$eig[1:i])/sum(pco$eig);
272 }
273 gofs1
274 gofs2
275 eigfracs
276 #R^2 for first four PCOs
277 corrs[1:4]*corrs[1:4]
278 #Plot eigenvalues
279 #(To save as PDF, uncomment next line)
280 pdf(file='EigenvaluesCulled80?.pdf', width=6, height=6);
281 pco <- cmdscale(d,eig=TRUE);
282 plot(1:length(pco$eig),pco$eig,pch=16,col="#00000080",bty="n",xlab="Eigenvalue number",ylab
        ="Eigenvalue",main="");
283 abline(h=0, col="grey");
284 dev.off();
285 ######
286
287
288 ###INFORMATION CONTAINED IN PCO AXES, BY CORRELATION OF DISTANCES###
289 #Run PCO with all eigenvectors returned
290 pco <- cmdscale(d, k=dim(d)[1]-1, eig=TRUE);
291 #There are this many positive eigenvectors (PCO axes)
292 valid <- length(pco$eig[pco$eig > 0]);
```

Code F.1: Morphospace.R (continued)

```
#Original distance matrix as "dist" type object (lower triangular matrix)
realdist <- as.dist(d);
#Squared
realdistsq <- realdist*realdist;
#Set up plot
#pdf(file='PCOAxis_Corr.pdf', width=5, height=50,bg="white");
#par(mfrow = c(10,1));
#Set up a vector to hold correlations
corrs <- vector(mode="numeric",length=valid);
#Now loop through through those PCO axes and plot
for (i in 1:valid)
{
  #Calculate pairwise distances (as Euclidean) on i PCO axes (in i-space)
  pcodist <- dist(pco$points[,1:i]);
  #Plot original distances against distances recovered by i PCO axes
  #plot(realdist,pcodist,pch=16,col="#80808020")
  pcodistsq <- pcodist*pcodist;
  #Now find correlation (Pearson's r) between the squared distances in d and those on PCO
  corrs[i] <- cor(realdistsq,pcodistsq);
}
#dev.off();
#Make plot of # of principal coordinates vs. correlation
pdf(file='PCO_distance_correlation_8080?.pdf',bg="white");
plot(corrs*corrs,log="y",pch=16,col="#0000080",bty="n",xaxs = "i",xlab="Principal
     coordinates",ylab="Squared correlation",xlim=c(0,sum(pco$eig > 0)+1));
axis(2, at=c(0,1), labels=c("",""), lwd.ticks=0)
dev.off();
######


###VARIANCE EXPLAINED BY PCO AXES (PUBLICATION 1)###
#Get the fonts ready
file.exists <- function( fname ) length(Sys.glob(fname))>0
absolute.path.to.font.files <-  "/Users/bkotrc/font/";
bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
## if you do not have the correct font types
for (i in 1:length(bera.names)) {
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".afm", sep="")) )
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".otf", sep="")) )
}
gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files,bera.names, ".afm", sep
     =""))
pdfname <- "varexplfig.pdf";
#Make a composite plot of previous plus this plot for publication
pdf(file=pdfname, bg="white", width=(7.2/cm(1)), height=(14.4/cm(1)), pointsize=7, family=
     gillsans);
#pdf(file=pdfname, bg="white", width=(7.2/cm(1)), height=(14.4/cm(1)), pointsize=7);
par(font.main=1, cex.main=1.5)
axthck <- 0.3;
par(mfrow=c(2,1), lwd=axthck);
plot(1:length(pco$eig),pco$eig,bty="n",type="n",axes=FALSE,xlab="Eigenvalue number",ylab="
     Eigenvalue",main="",lwd=0.5);
title(main="A",adj=1);
abline(h=0);
points(1:length(pco$eig),pco$eig,pch=3,lwd=0.8);
axis(1, lwd.ticks=axthck, lwd=0);
```

Code F.1: Morphospace.R (continued)

```
axis(2, lwd=axthck);
plot(1:length(corrs*corrs),corrs*corrs,log="y",axes=FALSE,type="n",pch=3,bty="n",main="",
     xaxs = "i",xlab="Principal coordinates",ylab="Squared correlation",xlim=c(0,sum(pco$
     eig > 0)+1));
title(main="B",adj=1);
axis(1, lwd=axthck);
axis(2, at=c(0,0.2,0.4,0.6,0.8,1), lwd=axthck);
points(1:length(corrs*corrs),corrs*corrs,pch=3,lwd=0.8);
dev.off();
#Now embed font in that file
embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font");
######


###PCO AXIS 'LOADINGS'/CHARACTER ASSOCIATION WITH AXES###
#Run PCO with all eigenvectors returned
pco <- cmdscale(d, k=dim(d)[1]-1, eig=TRUE);
#First, generate two matrices of size a x b, where
#a = number of characters, b = number of valid (positive) PCO axes
#One to contain the Cramér coefficient of association between
#characters and PCO axes
cramer <- matrix(0, nrow=dim(m)[2], ncol=dim(pco$points)[2]);
rownames(cramer) <- colnames(m);
#The other to contain the p-values of the Cramér coefficients
#(actually the p-values of the Pearson Chi-squared statistic)
pvals <- matrix(0, nrow=dim(m)[2], ncol=dim(pco$points)[2]);
rownames(pvals) <- colnames(m);
#Catch characters that have only one state that I may have missed
onestaters <- vector(length=ncol(m));
names(onestaters) <- colnames(m);
#Loop through each character
for(i in 1:ncol(m))
{
  #And loop through each PCO axis
  for(j in 1:ncol(pco$points))
  {
    #Set up data frame with two columns: 1) character states for character i
    #2) PCO score of PCO axis j
    ijdata <- as.data.frame(cbind(m[,i],as.numeric(pco$points[,j])));
    #Character states should be factors (discrete)
    ijdata[,1] <- as.factor(ijdata[,1]);
    #PCO scores should be continuous to start with
    ijdata[,2] <- as.numeric(ijdata[,2]);
    #Discretize PCO scores into 4 equal intervals
    ijdata[,2] <- cut(ijdata[,2], breaks=4);
    #Label columns of data frame
    colnames(ijdata) <- c("char_state","pco_score");
    #Take out invalid data
    ijdata <- ijdata[(ijdata[,1] != 'v' & ijdata[,1] != 'n' & ijdata[,1] != '?'),];
    #Now construct a contingency table for this data frame
    ijtable <- xtabs(~char_state + pco_score, data=ijdata, drop.unused.levels=TRUE);
    #This contingency table must be at least 2*2 to make sense
    #(can't measure association on a 2*1 table)
    if(sum(dim(ijtable) > 1) > 1)
    {
      #Save Cramér's V
      cramer[i,j] <- assocstats(ijtable)$cramer;
      #Save p-value
```

Code F.1: Morphospace.R (continued)

```
      pvals[i,j] <- assocstats(ijtable)$chisq_tests[2,3];
    }
    else #Table has only one column or row
    {
      cramer[i,j] <- NA;
      #Make pvals big so it gets thrown out as > 0.05 below
      pvals[i,j] <- 10;
      #One row?
      if(dim(ijtable)[1] == 1){onestaters[i] <- TRUE};
    }
  }
}
######


###PCO AXIS-CHARACTER ASSOCIATION PLOT (PUBLICATION 1)###
#Get the fonts ready
file.exists <- function( fname ) length(Sys.glob(fname))>0
absolute.path.to.font.files <-  "/Users/bkotrc/font/";
bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
## if you do not have the correct font types
for (i in 1:length(bera.names)) {
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".afm", sep="")) )
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".otf", sep="")) )
}
gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files,bera.names, ".afm", sep
    =""))
pdfname <- "loadings.pdf";
#Make a composite plot of previous plus this plot for publication
pdf(file=pdfname, bg="white", width=(14.8/cm(1)), height=(17.7/cm(1)), pointsize=7, family=
    gillsans);
par(font.main=1, cex.main=1.5)
axthck <- 0.3;
#Plot the results in a bubbleplot
par(fig=c(10/83,71/83,10/110,107/110));
par(mar=c(4,4,0.3,2));
#Define color palette
#my.col <- c(colorRampPalette(c("black","grey"))(5),colorRampPalette(c("black","grey"))(5))
    ;
my.col <- c(gray.colors(10, start = 0, end = 0.33, gamma = 2.2),gray.colors(90, start =
    0.33, end = 0.66, gamma = 2.2),gray.colors(402, start = 0.66, end = 1, gamma = 2.2));
txtsz <- 1
plot(1,1,bty="n",xaxs="i",yaxs="i",cex.main=txtsz,cex.lab=txtsz,cex.axis=txtsz,asp=1,xlim=c
    (0,ncol(pco$points)),ylim=c(ncol(m)+1,0),xlab="PCO Axis",ylab="Character",type="n",
    main="", lab=c(5,10,7),lwd=axthck);
legcex <- par("cex.axis");
abline(v=1:ncol(pco$points),col="gray",lwd=0.25);
abline(v=seq(10,60,by=10),col="gray",lwd=1);
abline(h=1:ncol(m),col="gray",lwd=0.25);
abline(h=seq(10,100,by=10),col="gray",lwd=1);
#Get rid of p-values greater than 0.05
p05 <- pvals;
p05[p05 > 0.05] <- NA;
cramerp05 <- cramer;
cramerp05[is.na(p05)] <- NA;
#Loop through characters
```

Code F.1: Morphospace.R (continued)

```
for(i in 1:ncol(m))
{
  if(sum(is.na(p05[i,])) < ncol(pco$points))
  {
    x <- (1:ncol(pco$points))[!(is.na(p05[i,]))];
    y <- (rep(i,ncol(pco$points))[!(is.na(p05[i,]))]);
    symbols(x=x, y=y, circles=cramerp05[i,!(is.na(p05[i,]))], bg=my.col[ceiling((p05[i,!is.
        na(p05[i,])])*10000)], fg="grey15", lwd=0.25, inches=FALSE, add=TRUE);
    #Highlight the ones with Cramér > 0.4:
    #x <- x[cramerp05[i,] > 0.4];
    #y <- y[cramerp05[i,] > 0.4];
    #symbols(x=x, y=y, circles=rep(0.8,length(y[cramerp05[i,] > 0.4])), bg="red", fg="grey
        ", lwd=0.5, inches=FALSE, add=TRUE);
  }
}
#Finally, add a legend for Cramér
par(fig=c(68/83,1,80/110,1),new=TRUE);
par(mar=c(0,0,0,0));
plot(1,1,type="n",xlim=c(-3,13),ylim=c(23,-3), xaxs="i",yaxs="i", axes=FALSE);
x <- 5;
txtal <- c(1,0.5);
text(x,0.5,labels="Cramér coefficient", xpd=TRUE, cex=legcex);
text(x,2,labels="(degree of association)", xpd=TRUE, cex=legcex*0.8, font=3);
symbols(c(x-2,x-2,x-2,x-2), c(4,4.9,6,7.7), circles=c(0.1,0.25,0.5,1), xpd=TRUE, inches=
    FALSE, bg=my.col[1], fg="grey15", lwd=0.25, add=TRUE);
text(c(x+2,x+2,x+2,x+2)+0.8, c(3.5,5,6.5,8)+0.25,labels=c("0.1","0.25","0.5","1"), xpd=TRUE
    , cex=legcex, adj=txtal);
#And for p-values
text(x,12.5,labels="p-value", xpd=TRUE, cex=legcex);
text(x,14,labels="(significance)", xpd=TRUE, cex=legcex*0.8, font=3);
symbols(c(x-2,x-2,x-2,x-2), c(16,17.5,19,20.5), circles=c(0.7,0.7,0.7,0.7), xpd=TRUE,
    inches=FALSE, bg=my.col[ceiling(c(0.05,0.01,0.001,0.0001)*10000)], fg="grey15", lwd
    =0.25, add=TRUE);
text(c(x+2,x+2,x+2,x+2)+0.8, c(16,17.5,19,20.5),labels=c("0.05","0.01","0.001","0"), xpd=
    TRUE, cex=legcex, adj=txtal);

#Add marginal sums as bar plots
xhistC <- colSums(cramerp05, na.rm=TRUE);
xhistP <- colMeans(p05, na.rm=TRUE);
yhistC <- rowSums(cramerp05, na.rm=TRUE);
yhistP <- rowMeans(p05, na.rm=TRUE);
#(mar=c(3.8,6,6.6,0));
par(fig=c(0,10/83,14.25/110,109.5/110),new=TRUE);
par(mar=c(0,1,0,0));
par(lwd=0.25);
barplot(rev(yhistC), axes=FALSE, offset=-10,space=0, horiz=TRUE, col=my.col[ceiling(yhistP*
    10000)], border="black", axisnames=FALSE);
#par(mar=c(0.5,4.5,0,17.2));
par(fig=c(15.5/83,69.75/83,0,10/110),new=TRUE);
par(mar=c(1,0,0,0));
par(lwd=0.25);
barplot(xhistC, axes=FALSE, space=0, border="black", col=my.col[ceiling(xhistP*10000)]);
dev.off();
#Now embed font in that file
embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font");
#Now find which characters have the highest association with axes 1, 2, and 3
#PCO1, from highest Cramér to lowest:
round(sort(cramerp05[,1],decreasing=TRUE),digits=2);
```

Code F.1: Morphospace.R (continued)

```
round(sort(cramerp05[,2],decreasing=TRUE),digits=2);
round(sort(cramerp05[,3],decreasing=TRUE),digits=2);
#From the whole matrix of values with p < 0.05, there are 27 with
#Cramer > 0.4
round(sort(cramerp05,decreasing=TRUE),digits=2);
######


###SEVEN CHARACTERS' PCO PLOT (PUBLICATION 1)###
#Seven individual characters' state distribution on the PCO axes 1-2
#Parameter bw controls black & white vs. color plotting
plotSevenPartFig(pco,mfull,bw=FALSE);
######


###PCO PLOT WITH PLOT SYMBOLS REFLECTING GROSS SHAPE (PUBLICATION 1)###
#PCO plot with shapes for plot symbols
#Get the fonts ready
file.exists <- function( fname ) length(Sys.glob(fname))>0
absolute.path.to.font.files <-  "/Users/bkotrc/font/";
bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
## if you do not have the correct font types
for (i in 1:length(bera.names)) {
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".afm", sep="")) )
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".otf", sep="")) )
}
gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files,bera.names, ".afm", sep
    =""))
pdfname <- "pcoshape.pdf";
#Make a composite plot of previous plus this plot for publication
pdf(file=pdfname, bg="white", width=(14.8/cm(1)), height=(10/cm(1)), pointsize=7, family=
    gillsans);
par(mar=c(5,5,1,15));
#Generate plot area
points <- pco$points;
#Green symbols version:
#symcol <- "#22990885";
#B&W version
symcol <- "#00000085";
axcol <- "black";
axthck <- 0.3;
ylab <- "PC 2";
xlab <- "PC 1";
plot(points,type="n",bty="n", asp=1,  main="",col='grey',xaxt="n",yaxt="n",xlab="",ylab="")
    ;
box(which = "plot", lty = "solid", col=axcol, lwd=axthck);
axis(side=1,at=c(-0.1,0,0.1),line=0.5, col=axcol, lwd=axthck);
mtext(xlab,side=1,line=3);
axis(side=2,at=c(-0.1,0,0.1,0.2),line=0.5, col=axcol, lwd=axthck);
mtext(ylab,side=2,line=3);
abline(h=0, lty=1, col=axcol, lwd=axthck);
abline(v=0, lty=1, col=axcol, lwd=axthck);
#Set plot symbol size
size <- 0.000025;
for (i in 1:nrow(m))
{
```

Code F.1: Morphospace.R (continued)

```
  drawShape(points[i,1],points[i,2],size,as.numeric(m[i,"X1"]),as.numeric(m[i,"X2"]),as.
      numeric(m[i,"X112"]),color=symcol);
}
#Key
x <- 0.255;
y <- 0.2;
sp <- 0.02;
xsp <- 0.015;
text(x-0.01,y+sp,"Outline shape in valve view (1)",adj=0,xpd=TRUE);
for(i in 0:2) drawShape(x,y-(i*sp),size,i,4,0,color=symcol,xpd=TRUE);
i <- 3;
drawShape(x,y-(i*sp),size,i,2,0,color=symcol,xpd=TRUE);
i <- 4;
drawShape(x,y-(i*sp),size,i,4,0,color=symcol,xpd=TRUE);
labs <- c("Elliptical","Rectangular","Rhombic","Ovate","Triangular");
for(i in 0:4) text(x+xsp,y-(i*sp),labs[i+1],adj=0,,font=3,cex=0.8,xpd=TRUE);
text(x-0.01,y-6*sp,"Aspect ratio in valve view (2)",adj=0,xpd=TRUE);
i <- 0;
drawShape(x-0.005,(y-7.3*sp)-(i*sp),size,0,i,0,color=symcol,xpd=TRUE);
i <- 1;
drawShape(x+0.005,(y-7*sp)-(i*sp),size,0,i,0,color=symcol,xpd=TRUE);
for(i in 2:4) drawShape(x,(y-7*sp)-(i*sp),size,0,i,0,color=symcol,xpd=TRUE);
for(i in 5:7) drawShape(x,(y-7*sp)-(i*sp),size,3,i,0,color=symcol,xpd=TRUE);
labs <- c("Linearis","Anguste","Anguste late","Late","1:1","Latissime","Depresse","
    Perdepresse");
i <- 0;
text(x+xsp,(y-7*sp)-(i*sp),labs[i+1],adj=0,,font=3,cex=0.8,xpd=TRUE);
i <- 1;
text(x+xsp,(y-7*sp)-(i*sp),labs[i+1],adj=0,,font=3,cex=0.8,xpd=TRUE);
for(i in 2:7) text(x+xsp,(y-7*sp)-(i*sp),labs[i+1],adj=0,,font=3,cex=0.8,xpd=TRUE);
text(x-0.01,y-16*sp,"Raphe (90)",adj=0,xpd=TRUE);
drawShape(x-0.0025,(y-17*sp),size,0,1,0,color=symcol,xpd=TRUE);
drawShape(x+0.0025,(y-18*sp),size,0,1,1,color=symcol,xpd=TRUE);
labs <- c("Raphe absent","Raphe present");
for(i in 0:1) text(x+xsp,(y-17*sp)-(i*sp),labs[i+1],adj=0,,font=3,cex=0.8,xpd=TRUE);
dev.off();
#Now embed font in that file
embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font");
######


###SENSITIVITY OF RESULTS TO ORDINATION PLOT (PUBLICATION 2)###
#Testing results to sensitivity of ordination procedure by
#comparing results under PCO to those under NMDS
plotOrdinationComparison(bins, binnames, samplingmode="in-bin", N, d)
######


###IMAGE-ANNOTATED PCO PLOT (PUBLICATION 1)###
#PUBLICATION FIGURE 2
#Images should be stored in a subfolder of the working directory named "Images"
#Specify a list of names to plot (also filenames of images)
name <- c("Planktoniella","Triceratium","Azpeitia","Paralia","Arachnoidiscus","
    Stephanodiscus","Cymatopleura","Fragilaria","Cymbella","Mastogloia","Navicula","
    Rhizosolenia","Hemiaulus","Odontella","Stephanopyxis","Biddulphia");
#Call plotting function
plotMS2DImages(pco$points,name,savePDF=TRUE);
######
```

Code F.1: Morphospace.R (continued)

```
614
615
616 ###PLOT MORPHOLOGICAL VS. MOLECULAR DISTANCE (PUBLICATION 1)###
617 #Get the fonts ready
618 file.exists <- function( fname ) length(Sys.glob(fname))>0
619 absolute.path.to.font.files <-  "/Users/bkotrc/font/";
620 bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
621 ## if you do not have the correct font types
622 for (i in 1:length(bera.names)) {
623   stopifnot( file.exists(paste(absolute.path.to.font.files,
624                                bera.names[i], ".afm", sep="")) )
625   stopifnot( file.exists(paste(absolute.path.to.font.files,
626                                bera.names[i], ".otf", sep="")) )
627 }
628 gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files,bera.names, ".afm", sep
      ="""))
629 pdfname <- "morphmoldists.pdf";
630 pdf(file=pdfname, bg="white", width=(7.2/cm(1)), height=(14.4/cm(1)), pointsize=7, family=
      gillsans);
631 par(font.main=1, cex.main=1.5)
632 axthck <- 0.3;
633 par(mfrow=c(2,1), lwd=axthck);
634 #Calculating the *patristic* distances, using the tree
635 #file supplied by Ulf Sörhannus
636 tr <- read.tree("ulf.nwk");
637 #Calculate patristic distance matrix
638 pdist <- cophenetic(tr);
639 #Sort this by rownames and colnames
640 pdist <- pdist[order(rownames(pdist)),];
641 pdist <- pdist[,order(colnames(pdist))];
642 #Get a submatrix with only the first species of each genus
643 dnames <- rownames(pdist);
644 splits <- strsplit(dnames,split="_");
645 splits <- unlist(strsplit(dnames,split="_"));
646 gennames <- matrix(splits, nrow=length(dnames), ncol=2, byrow=TRUE)[,1];
647 firsts <- vector(length=length(gennames));
648 firsts[1] <- TRUE;
649 for(j in 2:length(gennames)){
650   if(gennames[j] != gennames[j-1]){
651     firsts[j] <- TRUE;
652   }
653 }
654 pdistgen <- pdist[firsts,firsts];
655 rownames(pdistgen) <- unique(gennames);
656 colnames(pdistgen) <- unique(gennames);
657 #Now find the set of genus names that molecular and morphological data share
658 shared_gen <- intersect(rownames(d),gennames);
659 #Matrices of shared genera
660 #Distance matrices of shared genera
661 d_morph <- d[shared_gen,shared_gen];
662 d_mol <- pdistgen[shared_gen,shared_gen];
663 #Now plot
664 plot(as.dist(d_mol),as.dist(d_morph), xaxs = "i", axes=FALSE, xpd=NA, yaxs = "i",col="
      #00000050",bty="n",xlab="Patristic molecular distance",ylab="Morphological distance",
      pch=16, xlim=c(0,max(as.dist(d_mol))), ylim=c(0,max(as.dist(d_mol))));
665 title(main="A",adj=1);
666 axis(1, lwd=axthck, at=c(0,0.2,0.4,0.6));
667 axis(2, lwd=axthck, at=c(0,0.2,0.4,0.6));
```

Code F.1: Morphospace.R (continued)

```
#Linear regression:
reg1 <- lm(as.dist(d_morph)~as.dist(d_mol));
#Calculate distances among sequences directly
#Read in Ulf Sörhannus' sequence alignments from his 2007 paper
#in Micropaleontology
rna <- read.alignment(file="ulf.aln",format="clustal")
d_mol <- as.matrix(dist.alignment(rna, matrix="identity"));
#Sort this by rownames and colnames
d_mol <- d_mol[order(rownames(d_mol)),];
d_mol <- d_mol[,order(colnames(d_mol))];
#Now we need to fix the rows and columns of the molecular distance matrix
#so that it's species rather than genus, and that the name is genus only
#so that it can be matched against the morphological matrix
dnames <- rownames(d_mol);
splits <- strsplit(dnames,split="_");
splits <- unlist(strsplit(dnames,split="_"));
gennames <- matrix(splits, nrow=length(dnames), ncol=2, byrow=TRUE)[,1];
firsts <- vector(length=length(gennames));
firsts[1] <- TRUE;
for(j in 2:length(gennames)){
  if(gennames[j] != gennames[j-1]){
    firsts[j] <- TRUE;
  }
}
#The molecular distance matrix of genera only (using the first listed species
#as representative for the genus)
d_mol_genus <- d_mol[firsts,firsts];
colnames(d_mol_genus) <- unique(gennames);
rownames(d_mol_genus) <- unique(gennames);
#Now find the set of genus names that molecular and morphological data share
shared_gen <- intersect(rownames(d),gennames);
#Distance matrices of shared genera
d_morph <- d[shared_gen,shared_gen];
d_mol <- d_mol_genus[shared_gen,shared_gen];
#Plot to inspect
plot(as.dist(d_mol),as.dist(d_morph),bty="n",xpd=TRUE,xaxs="i",yaxs="i",axes=FALSE,xlab="
     Raw molecular distance",ylab="Morphological distance", pch=16, col="#00000050",xlim=c
     (0,max(as.dist(d_morph))), ylim=c(0,max(as.dist(d_morph))));
axis(1, lwd=axthck);
axis(2, lwd=axthck);
title(main="B",adj=1);
#Linear regression:
reg1 <- lm(as.dist(d_morph)~as.dist(d_mol));
#Calculate r^2
#rsq[i] <- cor(as.dist(d_mol),as.dist(d_morph),method="pearson")^2;
#text(0.2,0.05, paste("Morph. dist. = ", round(reg1$coefficients[1],2), "+ ", round(reg1$
     coefficients[2],2), "(Mol. dist.),    R2 = ",signif(rsq[i],digits=3)));
#Plot up the r^2 values
#plot(rsq, xlab="Index of distance algorithm", ylab=expression(italic(R^2)));
#text(rsq,labels=gdmodel,pos=4);
#Mantel test: raw sequences
#mant <- mantel.test(d_mol,d_morph,nperm=1000000,alternative="two.sided");
dev.off();
#Now embed font in that file
embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font");
#Mantel test: patristic
#mant <- mantel.test(d_mol,d_morph,nperm=1000000,alternative="two.sided");
######
```

Code F.1: Morphospace.R (continued)

```
723
724
725 ###PLOT PCO COMPARED TO PHYLOGENY (PUBLICATION 1)###
726 #Read in Ulf's tree
727 tr <- read.tree("ulf.nwk");
728 #Define list of species to plot on tree
729 #(this is one species from each genus shared between the morphological
730 #and molecular data, chosen to best represent the position of the genus
731 #in the tree)
732 splist <- c("Paralia_sol", "Stephanopyxis_palmeriana","Podosira_stelligera","Melosira_
      octogona","Aulacoseira_subarctica","Rhizosolenia_setigeraB","Corethron_criophilum","
      Stellarima_microtrias","Coscinodiscus_granii","Actinoptychus_sinensis","Actinocyclus_
      curvatulus","Rhaphoneis_belgicae","Delphineis_sp","Thalassionema_sp","Synedra_sp","
      Fragilaria_sp","Rhabdonema_sp","Grammatophora_oceanicaAUST","Diatoma_tenue","
      Fragilariopsis_sublineata","Nitzschia_frustulum","Navicula_diserta","Pleurosigma_sp","
      Gyrosigma_sp","Amphora_montana","Surirella_fastuosa","Achnanthes_sp","Cocconeis_
      molesta","Gomphonema_pseudaugur","Eunotia_sp","Bacillaria_paxillifer","Lithodesmium_
      undulatum","Odontella_sinensis","Pleurosira_laevis","Biddulphia_sp","Cymatosira_
      belgica","Porosira_pseudodenticulata","Skeletonema_spA","Thalassiosira_eccentrica","
      Stephanodiscus_hantzschii","Cyclotella_scaldensis","Eucampia_antarctica","Chaetoceros_
      sp","Planktoniella_sol");
733 #Plot the tree with only those tips
734 trim <- drop.tip(tr, which(!tr$tip.label %in% splist));
735 #Add labels identifying the numbers for the internal nodes
736 nodelabels();
737 #Now rotate nodes to a sensible order
738 trim <- rotate(trim, 45);
739 trim <- rotate(trim, 46);
740 trim <- rotate(trim, 80);
741 trim <- rotate(trim, 60);
742 trim <- rotate(trim, 61);
743 trim <- rotate(trim, 62);
744 trim <- rotate(trim, 67);
745 trim <- rotate(trim, 64);
746 #Chop out the species names (just leave genus names)
747 dnames <- trim$tip.label;
748 splits <- strsplit(dnames,split="_");
749 splits <- unlist(strsplit(dnames,split="_"));
750 gennames <- matrix(splits, nrow=length(dnames), ncol=2, byrow=TRUE)[,1];
751 trim$tip.label <- gennames;
752 #Get the fonts ready
753 file.exists <- function( fname ) length(Sys.glob(fname))>0
754 absolute.path.to.font.files <-  "/Users/bkotrc/font/";
755 bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
756 ## if you do not have the correct font types
757 for (i in 1:length(bera.names)) {
758   stopifnot( file.exists(paste(absolute.path.to.font.files,
759                                bera.names[i], ".afm", sep="")) )
760   stopifnot( file.exists(paste(absolute.path.to.font.files,
761                                bera.names[i], ".otf", sep="")) )
762 }
763 gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files,bera.names, ".afm", sep
      =""))
764 pdfname <- "pcophylo.pdf";
765 #Plot it
766 pdf(file=pdfname, bg="white", width=(14.8/cm(1)), height=(21.5/cm(1)), pointsize=7, family=
      gillsans, colormodel="cmyk");
767 par(fig=c(0,(2.5/5),0,1));
```

Code F.1: Morphospace.R (continued)

```
par(mar=c(0,0,0,0));
track <- plot(trim, use.edge.length=FALSE, cex=1, label.offset=1.75);
#Add lines & labels to denote "clades" (actually, mostly paraphyletic groups)
x <- track$x.lim[2]+(track$x.lim[2])*0.025;
#Radial Centrics
ybot <- 1;
ytop <- 11;
segments(x,ybot,x,ytop,lwd=2);
text(x+2.5,ybot+((ytop-ybot)/2),"Radial Centrics",srt=270, cex=1.25, xpd=NA);
yloc1 <- grconvertY(y=ybot+((ytop-ybot)/2), from="user", to="ndc");
#Bi- & Multipolar Centrics, Thalassiosirales
ybot <- 12;
ytop <- 24;
segments(x,ybot,x,ytop,lwd=2);
text(x+2.5,ybot+((ytop-ybot)/2),"Bi- & Multipolar Centrics",srt=270, cex= 1.25, xpd=NA);
yloc2 <- grconvertY(y=ybot+((ytop-ybot)/2), from="user", to="ndc");
#Araphids
ybot <- 25;
ytop <- 32;
segments(x,ybot,x,ytop,lwd=2);
text(x+2.5,ybot+((ytop-ybot)/2),"Araphids",srt=270, cex= 1.25, xpd=NA);
yloc3 <- grconvertY(y=ybot+((ytop-ybot)/2), from="user", to="ndc");
#Raphids
ybot <- 33;
ytop <- 44;
segments(x,ybot,x,ytop,lwd=2);
text(x+2.5,ybot+((ytop-ybot)/2),"Raphids",srt=270, cex= 1.25, xpd=NA);
yloc4 <- grconvertY(y=ybot+((ytop-ybot)/2), from="user", to="ndc");
#Get device locations for dotted lines joining group to PCO
yl <- grconvertY(y=c(1,11,12,24,25,32,33,44), from="user", to="ndc");
xl <- grconvertX(x=x, from="user", to="ndc");
#Now set up a vector of colors to represent each genus (=tip)
tipcolors <- colors()[c
     (504,33,35,36,24,190,333,121,131,565,598,76,504,33,36,24,190,76,121,565,76,504,33,35,36,
     ];
#This vector is in the order of the tips as they appear, but the tips are numbered
     differently, so resort:
tiporder <- c
     (21,20,19,18,24,23,22,16,15,14,17,25,31,30,29,28,27,26,33,32,13,12,11,10,9,8,7,6,5,4,3,2
     ;
names(tipcolors) <- trim$tip.label[tiporder];
#Now plot up little colored circles for each genus on the phylogeny
tiplabels(tip=tiporder, col="black", pch=21, bg=tipcolors, lwd=0.3, adj=1, cex=1.5);
#Now add the PCO plots, same order as above
plotht <- 0.1;
plotw <- 2*plotht*(215/148);
par(fig=c((3/5),(3/5)+plotw,yloc1-plotht,yloc1+plotht),new=TRUE);
plotMS2Dgen(pco$points,tipcolors[34:44]);
yr <- grconvertY(y=par("usr")[3:4], from="user", to="ndc");
par(fig=c((3/5),(3/5)+plotw,yloc2-plotht,yloc2+plotht),new=TRUE);
plotMS2Dgen(pco$points,tipcolors[21:33]);
yr <- c(yr,grconvertY(y=par("usr")[3:4], from="user", to="ndc"));
par(fig=c((3/5),(3/5)+plotw,yloc3-plotht,yloc3+plotht),new=TRUE);
plotMS2Dgen(pco$points,tipcolors[13:20]);
yr <- c(yr,grconvertY(y=par("usr")[3:4], from="user", to="ndc"));
par(fig=c((3/5),(3/5)+plotw,yloc4-plotht,yloc4+plotht),new=TRUE);
plotMS2Dgen(pco$points,tipcolors[1:12]);
yr <- c(yr,grconvertY(y=par("usr")[3:4], from="user", to="ndc"));
```

Code F.1: Morphospace.R (continued)

```
xr <- grconvertX(x=rep(par("usr")[1],8), from="user", to="ndc");
#Now add dashed lines linking PCO to groups
par(fig=c(0,1,0,1), new=TRUE);
plot(0, 0, type="n", axes=FALSE, bty="n");
xl <- grconvertX(x=xl, from="ndc", to="user");
xr <- grconvertX(x=xr, from="ndc", to="user");
yr <- grconvertY(y=yr, from="ndc", to="user");
yl <- grconvertY(y=yl, from="ndc", to="user");
segments(x0=xl, x1=xr, y0=yl, y1=yr, lwd=0.3, lty="dashed", col="black");
dev.off();
#Now embed font in that file
embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font");
######


###PREP NEPTUNE AND SET TIME BINS###
#Prepare Neptune –database-note this only needs to be once, to the raw Neptune
#database as downloaded, subsequent runs can use N as loaded from NeptuneProcessed.txt
N <- prepNeptune(N_raw);
#Make sure there are no entries in the matrix that aren't
#also in Neptune
setdiff(row.names(m),unique(N$Genus));
#And the converse, throw out Neptune occurrences that aren't in the matrix
N <- N[N$Genus %in% row.names(d),];
######


###PLOT PCO STACKED THROUGH TIME (PUBLICATION 1)###
#Run PCO with all eigenvectors returned
pco <- cmdscale(d, k=dim(d)[1]-1, eig=TRUE);
#Plot
plotMS3D(pco$points[,1:2], msls, savePDF=TRUE, mode="color by bin");
######


###PLOT PCO STACKED, SHOWING OCCURRENCES (PUBLICATION 2)####
#Plot with plot symbol size reflecting number of occurrences
#We are only interested in the Neptune data in this second paper
#since we need occurrence data to perform subsampling, so we need
#to reset the time bins to include the Cenozoic only.
#Set time boundaries
bins <- c(54.5,33.7,23.8,5.32,1.85);
#Get rid of pre-Cenozoic occurrences
N <- N[N$Sample.Age < 65,];
#Set time bin names
binnames <- c("Paleocene","Eocene","Oligocene","Miocene","Pliocene","Pleistocene");
#Get lists of genus names for each time bin (mode can be "in-bin" or "range-through")
msls <- getBinLists(N, bins, binnames, mode="range-through");
#Get occurrences for each bin and each genus
#Loop through time bins
binbounds <- c(65,bins,0);
sizes <- as.list(1);
meanoccs <- vector(mode="numeric", length=length(msls));
for(i in 1:length(msls))
{
  #Get vector ready
  sizes[[i]] <- vector(mode="numeric", length=length(msls[[i]]));
  #Neptune for the time bin i
```

Code F.1: Morphospace.R (continued)

```
  Nbin <-  N[N$Sample.Age < binbounds[i] & N$Sample.Age >= binbounds[i+1], ];
  #Loop through each genus in bin i
  for(j in 1:length(msls[[i]]))
  {
    #Count occurrences for the genus
    sizes[[i]][j] <- length(Nbin$Genus[Nbin$Genus == msls[[i]][j]]);
  }
  #Store the mean number of occurrences (for use in labeling plots)
  meanoccs[i] <- mean(sizes[[i]]);
  #Standardize the sizes to the mean number of occurrences
  sizes[[i]] <- sizes[[i]]/mean(sizes[[i]]);
  #Take square root of sizes
  sizes[[i]] <- sqrt(sizes[[i]]);
  #Make smallest dots not too small
  sizes[[i]][sizes[[i]] < 0.2] <- 0.2;
}
sizes$meanoccs <- meanoccs;
plotMS3D(pco$points[,1:2], msls, savePDF=TRUE, mode="color by bin", sizes=sizes);
######


###ALPHA DISPARITY PLOT (PUBLICATION 2)###
#"Alpha disparity—"the average disparity per-list (i.e. per borehole)
#Calculated for mean pairwise distance and alpha shape volume
#Now featuring confidence intervals!
#Diversity and disparity metrics through time
#Run PCO with all eigenvectors returned
pco <- cmdscale(d, k=dim(d)[1]-1, eig=TRUE);
#Make sure there are no entries in the matrix that aren't
#also in Neptune (actually, it's probably OK for those to be in, since
#only the relevant taxa will be selected for calculations by getBinLists())
setdiff(row.names(m),unique(N$Genus));
#And the converse, throw out Neptune occurrences that aren't in the matrix
#(this step is crucial or it will break)
N <- N[N$Genus %in% row.names(d),];
#Time bins, taxon lists in bins:
#Reset time bin boundaries to 2-myr intervals, preceded by E and L Cretaceous bins
#(N.b.: because 65 myrs is an odd number, the youngest interval is only 1 myr long)
bins <- c(seq(from=62,to=2,by=-2));
#Set time bin names
binnames <- c(seq(from=64,to=2,by=-2)-1);
#Make sure most recent version of function is in memory
source("MorphospaceFunctions.R");
#Call functions to calculate "alpha disparity" in two ways and plot
alphavols <- getAlphaDispVol(bins, binnames, N, d);
alphadists <- getAlphaDispMPWD(bins, binnames, N, d);
plotAlphaDisparity(alphavols,alphadists);
######


###DIVERSITY AND DISPARITY PLOTS (PUBLICATION 1)###
#Diversity and disparity metrics through time
#Time bins, taxon lists in bins:
#Reset time bin boundaries to 2-myr intervals, preceded by E and L Cretaceous bins
#(N.b.: because 65 myrs is an odd number, the youngest interval is only 1 myr long)
bins <- c(99.6, 64, seq(from=62,to=2,by=-2));
#Set time bin names
binnames <- c(112, 73.825, seq(from=64,to=2,by=-2)-1);
```

Code F.1: Morphospace.R (continued)

```
#Which taxon sampling mode to use?
#Options: ”in-bin”, ”range-through”,”uw”,”cr”
samplingmode <- ”range-through”;
#Call function to calculate disparity & diversity metrics and plot them to file
#named ”dispdiv.pdf”
disp <- plotDivDispPubLg(bins, binnames, samplingmode, N, d, sptrials=200,gentrials=100);
######


###DIVERSITY AND DISPARITY PLOTS (PUBLICATION 2)###
#Diversity and disparity metrics through time
#Run PCO with all eigenvectors returned
pco <- cmdscale(d, k=dim(d)[1]-1, eig=TRUE);
#Make sure there are no entries in the matrix that aren't
#also in Neptune (actually, it's probably OK for those to be in, since
#only the relevant taxa will be selected for calculations by getBinLists())
setdiff(row.names(m),unique(N$Genus));
#And the converse, throw out Neptune occurrences that aren't in the matrix
#(this step is crucial or it will break)
N <- N[N$Genus %in% row.names(d),];
#Time bins, taxon lists in bins:
#Reset time bin boundaries to 2-myr intervals, preceded by E and L Cretaceous bins
#(N.b.: because 65 myrs is an odd number, the youngest interval is only 1 myr long)
bins <- c(seq(from=62,to=2,by=-2));
#Set time bin names
binnames <- c(seq(from=64,to=2,by=-2)-1);
#Which taxon sampling mode to use?
#Options: ”in-bin”, ”range-through”,”uw”,”cr”,”sqs”
samplingmode <- ”cr”;
#Call function to calculate disparity & diversity metrics and plot them to file
#named ”dispdiv.pdf”
disp <- plotDivDispPub2Lg(bins, binnames, samplingmode, N, d, sptrials=10000,gentrials
    =10000);
######


###REALIZED CHARACTER STATES PLOT (PUBLICATION 1)###
#Array of all the character states in the morphospace
#List of morphospace character names
characters <- colnames(m);
#Invalid states
inv <- c(”?”,”n”,”v”);
#Character states for each character (valid states only)
charstates <- paste(characters[1],setdiff(sort(unique(m[[characters[1]]])),inv),sep=”:”)
for(i in 2:length(colnames(m)))
{
  charstates <- c(charstates,paste(characters[i],setdiff(sort(unique(m[[characters[i]]])),
      inv),sep=”:”))
}
#Create equivalent of m matrix resolved into binary (boolean) states
mstates <- matrix(data=FALSE, nrow=nrow(d), ncol=length(charstates));
colnames(mstates) <- charstates;
rownames(mstates) <- rownames(m);
#Loop through the whole new matrix element by element
#Rows
for(i in 1:nrow(mstates))
{
  #Columns (character states)
```

Code F.1: Morphospace.R (continued)

```
  for (j in 1:ncol(mstates))
  {
    ch <- unlist(strsplit(charstates[j],split=":"))[1];
    st <- unlist(strsplit(charstates[j],split=":"))[2];
    if(m[i,ch] == st)
    {
      mstates[i,j] <- TRUE;
    }
  }
}
#Make sure there are no entries in the matrix that aren't
#also in Neptune
setdiff(row.names(m),unique(N$Genus));
#And the converse, throw out Neptune occurrences that aren't in the matrix
N <- N[N$Genus %in% row.names(d),];
#Time bins, taxon lists in bins:
#Reset time bin boundaries to 2-myr intervals, preceded by E and L Cretaceous bins
#(N.b.: because 65 myrs is an odd number, the youngest interval is only 1 myr long)
bins <- c(99.6, 64, seq(from=62,to=2,by=-2));
#Set time bin names
binnames <- c(112, 73.825, seq(from=64,to=2,by=-2)-1);
#Get taxon lists
msls <- getBinLists(N, bins, binnames, mode="range-through");
#Variable to hold # of realized character states
nrstates <- vector(mode="numeric", length=length(msls));
#And number of genera for each bin
ngenera <- vector(mode="numeric", length=length(msls));
#Loop through each time bin
for(i in 1:length(msls))
{
  #Set up list of character states for time bin
  charlist <- NULL;
  #In each time bin, loop through all the taxa in that bin
  for(j in 1:length(msls[[i]]))
  {
    #List of realized character states for this taxon
    charlist <- union(charlist,colnames(mstates)[mstates[msls[[i]][j],]]);
  }
  #Now determine length of list of collected character states
  nrstates[i] <- length(unique(charlist));
  ngenera[i] <- length(msls[[i]]);
}
#Get the fonts ready
file.exists <- function( fname ) length(Sys.glob(fname))>0
absolute.path.to.font.files <-  "/Users/bkotrc/font/";
bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
## if you do not have the correct font types
for (i in 1:length(bera.names)) {
  stopifnot( file.exists(paste(absolute.path.to.font.files, bera.names[i], ".afm", sep="")))
  stopifnot( file.exists(paste(absolute.path.to.font.files, bera.names[i], ".otf", sep="")))
}
gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files, bera.names, ".afm",
    sep=""))
pdfname <- "realstates.pdf";
#Open PDF device
```

Code F.1: Morphospace.R (continued)

```
pdf(file=pdfname, bg="white", width=(14.8/cm(1)), height=(7.8/cm(1)), pointsize=7, family=gillsans, colormodel="cmyk");
#Now plot it up
par(mfrow=c(2,1), mar=c(1.5,4.1,1.5,4.1), oma=c(5.1,7.5,0,7.5));
axthck <- 0.3;
y <- nrstates;
plot(names(msls),y,pch=16,cex=0.8,axes=FALSE,bty="n",xpd=TRUE,xlim=c(112,0),ylim=c((min(y,na.rm=TRUE)-(diff(range(y,na.rm=TRUE))*0.19)),max(y)),xlab="",ylab="Realized character states");
axis(1,lwd=axthck)
axis(2,lwd=axthck)
berg95(line=axthck);
points(names(msls)[3:length(msls)],y[3:length(y)],type="l");
title(main="A",adj=0.05,line=-1, cex.main=1.5, font.main=1);
y <- nrstates/ngenera;
plot(names(msls),y,pch=16,cex=0.8,axes=FALSE,bty="n",xpd=TRUE,xlim=c(112,0),ylim=c((min(y,na.rm=TRUE)-(diff(range(y,na.rm=TRUE))*0.19)),max(y)),xlab="",ylab="Realized states per genus");
mtext(text="Geologic Time (Ma)",side=1,line=3,xpd=TRUE,cex=par("cex"));
axis(1,lwd=axthck)
axis(2,lwd=axthck)
berg95(line=axthck);
points(names(msls)[3:length(msls)],y[3:length(y)],type="l");
title(main="B",adj=0.05,line=-1, cex.main=1.5, font.main=1);
#Close PDF device
dev.off();
#Now embed font in that file
embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font");
######


###NEPTUNE SAMPLING INTENSITY PLOT (PLANKTONTECH CHAPTER)###
#Number of Neptune lists through time
#Culled to include only genera in the morphospace analysis
#And the converse, throw out Neptune occurrences that aren't in the matrix
N <- N[N$Genus %in% row.names(d),];
#Calculate number of lists
source("DiversityFunctions.R");
Nstats <- getSamplingIntensity(N=N,bins=32,agemax=64,agemin=0);
y <- Nstats$lists;
#Get the fonts ready
file.exists <- function( fname ) length(Sys.glob(fname))>0
absolute.path.to.font.files <-  "/Users/bkotrc/font/";
bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
## if you do not have the correct font types
for (i in 1:length(bera.names)) {
  stopifnot( file.exists(paste(absolute.path.to.font.files, bera.names[i], ".afm", sep="")) )
  stopifnot( file.exists(paste(absolute.path.to.font.files, bera.names[i], ".otf", sep="")) )
}
gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files, bera.names, ".afm", sep=""))
pdfname <- "neptsamp.pdf";
#Open PDF device
pdf(file=pdfname, bg="white", width=(7.2/cm(1)), height=(3.65/cm(1)), pointsize=7, family=gillsans, colormodel="cmyk");
#Now plot it up
```

Code F.1: Morphospace.R (continued)

```
par(mar=c(5.1,4.1,0.1,1.1));
axthck <- 0.3;
plot(Nstats$midpoint,y,pch=16,cex=0.8,axes=FALSE,bty="n",xpd=TRUE,xlim=c(112,0),ylim=c((min(y,na.rm=TRUE)-(diff(range(y,na.rm=TRUE))*0.19)),max(y)),xlab="",ylab="Number of lists");
mtext(text="Geologic Time (Ma)",side=1,line=3,xpd=TRUE,cex=par("cex"));
axis(1,lwd=axthck)
axis(2,lwd=axthck)
berg95(line=axthck);
#Close PDF device
dev.off();
#Now embed font in that file
embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font");
######


###PLOT OF CHARACTER SETS THROUGH TIME (PUBLICATION 2)###
#Be sure to run the above block of code first for character state matrix mstates
#We are only going to do this for the Neptune data
bins <- c(seq(from=62,to=2,by=-2));
#Set time bin names
binnames <- c(seq(from=64,to=2,by=-2)-1);
#Get taxon lists
msls <- getBinLists(N, bins, binnames, mode="range-through");

#Open a file to plot the results
#Get the fonts ready
  file.exists <- function( fname ) length(Sys.glob(fname))>0
  absolute.path.to.font.files <-  "/Users/bkotrc/font/";
  bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
  ## if you do not have the correct font types
  for (i in 1:length(bera.names)) {
    stopifnot( file.exists(paste(absolute.path.to.font.files, bera.names[i], ".afm", sep="")) )
    stopifnot( file.exists(paste(absolute.path.to.font.files, bera.names[i], ".otf", sep="")) )
  }
  gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files, bera.names, ".afm", sep=""))
  pdfname <- "charpercents.pdf";
  #Open PDF device
  pdf(file=pdfname, bg="white", width=(7.2/cm(1)), height=(13/cm(1)), pointsize=7, family=gillsans, colormodel="cmyk");

par(las=1, mfrow=c(4,1), mar=c(1.15,4.5,0,4.1), oma=c(5.1,0,1,0));
#Set up character states of interest for the first set of characters
#("Predation resistance")
#Character states of interest
chintname <- "Suggesting more predation resistance";
chint <- list("X1:0", #Elliptical outline
              "X2:4", #1:1 aspect ratio
              "X12:2", #Convex valve face
              "X27:2", #Numerous elongated or ornamented spines or conspicuous marginal processes
              "X37:1", #Mantle spines
              c("X43:1","X43:2","X43:3"), #External costae/ribs
              c("X86:1","X86:2"), #Ann. pseudoseptum
              c("X87:1","X87:2"), #Pseudoseptae
```

Code F.1: Morphospace.R (continued)

```
             c("X88:1"), #Pseudoseptae crossbars
             c("X89:1","X89:2"), #Parallel pseudoseptae
             "X90:1", #Marginal ribs
             "X121:1", #Raphe keel
             "X122:1" #Fibulae
             );
#Set other values of those characters
chothname <- "Suggesting less predation resistance";
choth  <- list(c("X1:1","X1:2","X1:3","X1:4"),
    c("X2:0","X2:1","X2:2","X2:3","X2:6","X2:7"),
    c("X12:0","X12:1","X12:3","X12:4","X12:5"),
    c("X27:0","X27:1"),
    c("X37:0"),
    c("X43:0","X43:4","X43:5"),
    c("X86:0"),
    c("X87:0"),
    c("X88:0"),
    "X89:0",
    c("X90:0"),
    "X121:0",
    "X122:0"
    );
#Make sure chint and choth are the same length!
if(length(chint)==length(choth)) print("You're OK") else print("Length mismatch!")
#Get percentages of genera having first character of interest for each time bin
pctchint <- getCharPct(chint[[1]], choth[[1]], mstates, msls);
#Now loop over the remaining characters of interest
for(i in 2:length(chint))
{
  #Get percentages of genera having character of interest for each time bin
  pctchint <- rbind(pctchint,getCharPct(chint[[i]], choth[[i]], mstates, msls));
}
#If any values are NaN, turn them to NA so the colMeans will ignore them
if(length(chint)==length(choth)) print("You're OK") else print("Length mismatch!")
#Get the average percentages over the set of characters examined
pctchint <- colMeans(pctchint, na.rm=TRUE);
#Plot it
plotCharPct(pctchint,binnames,chintname,chothname);
title(main="A",adj=0.05,line=-2, cex.main=1.5, font.main=1);
#Set up character states of interest for the second set of characters
#("Linkage")
#Character states of interest
chintname <- "Suggesting cell-cell linkage present";
chint <- list(c("X18:1","X18:2","X18:3"),
             c("X27:1","X27:2"),
             "X28:1",
             c("X55:1","X55:2"),
             c("X59:1","X59:2"),
             c("X96:1","X96:2"),
             "X98:3"
             );
#Set other values of those characters
chothname <- "Suggesting cell-cell linkage absent";
choth  <- list("X18:0",
              "X27:0",
              "X28:0",
              "X55:0",
              "X59:0",
```

Code F.1: Morphospace.R (continued)

```
              "X96:0",
              "X98:0"
              );
#Make sure chint and choth are the same length!
if(length(chint)==length(choth)) print("You're OK") else print("Length mismatch!")
#Get percentages of genera having first character of interest for each time bin
pctchint <- getCharPct(chint[[1]], choth[[1]], mstates, msls);
#Now loop over the remaining characters of interest
for(i in 2:length(chint))
{
  #Get percentages of genera having character of interest for each time bin
  pctchint <- rbind(pctchint,getCharPct(chint[[i]], choth[[i]], mstates, msls));
}
#If any values are NaN, turn them to NA so the colMeans will ignore them
#pctchint[is.na(pctchint)] <-  NA;
#Get the average percentages over the set of characters examined
pctchint <- colMeans(pctchint, na.rm=TRUE);
#Plot it
plotCharPct(pctchint,binnames,chintname,chothname,bty="u");
title(main="B",adj=0.05,line=-2, cex.main=1.5, font.main=1);
#Set up character states of interest for the third set of characters
#("Antivirus")
#Character states of interest
chintname <- "Suggesting more protection against viral attack";
chint <- list(c("X41:1","X41:2","X41:3","X41:4"),
              "X68:0",
              "X69:5",
              "X70:2"
              );
#Set other values of those characters
chothname <- "Suggesting less protection against viral attack";
choth  <- list("X41:0",
               c("X68:1","X68:2","X68:3"),
               c("X69:0","X69:1","X69:2","X69:3","X69:4"),
               c("X70:0","X70:1")
               );
#Make sure chint and choth are the same length!
if(length(chint)==length(choth)) print("You're OK") else print("Length mismatch!")
#Get percentages of genera having first character of interest for each time bin
pctchint <- getCharPct(chint[[1]], choth[[1]], mstates, msls);
#Now loop over the remaining characters of interest
for(i in 2:length(chint))
{
  #Get percentages of genera having character of interest for each time bin
  pctchint <- rbind(pctchint,getCharPct(chint[[i]], choth[[i]], mstates, msls));
}
#If any values are NaN, turn them to NA so the colMeans will ignore them
#pctchint[is.na(pctchint)] <-  NA;
#Get the average percentages over the set of characters examined
pctchint <- colMeans(pctchint, na.rm=TRUE);
#Plot it
plotCharPct(pctchint,binnames,chintname,chothname,bty="u");
title(main="C",adj=0.05,line=-2, cex.main=1.5, font.main=1);
#Set up character states of interest for the fourth set of characters
#("Silica Use")
#Character states of interest (less silica)
chintname <- "Suggesting more silica use";
chint <- list("X61:0",
```

Code F.1: Morphospace.R (continued)

```
              "X59:0",
              "X70:0",
              "X80:1",
              "X98:3",
              "X27:2",
              "X29:0",
              "X35:0",
              "X36:0",
              "X37:0",
              c("X38:0","X38:2","X38:3"),
              "X43:0",
              "X49:0",
              c("X53:0","X53:1"),
              "X55:0",
              "X56:0",
              "X83:0",
              "X84:0"
              );
#Set other values of those characters (more silica)
chothname <- "Suggesting less silica use";
choth  <- list(c("X61:1","X61:2","X61:3"),
               c("X59:1","X59:2"),
               c("X70:1","X70:2"),
               "X80:1",
               "X98:0",
               c("X27:0","X27:1"),
               c("X29:1","X29:2","X29:3"),
               "X35:1",
               "X36:1",
               "X37:1",
               "X38:5",
               c("X43:1","X43:2","X43:3","X43:4","X38:5"),
               "X49:1",
               c("X53:2","X53:4"),
               c("X55:1","X55:2"),
               "X56:1",
               "X83:1",
               "X84:1"
               );
#Make sure chint and choth are the same length!
if(length(chint)==length(choth)) print("You're OK") else print("Length mismatch!")
#Get percentages of genera having first character of interest for each time bin
pctchint <- getCharPct(chint[[1]], choth[[1]], mstates, msls);
#Now loop over the remaining characters of interest
for(i in 2:length(chint))
{
  #Get percentages of genera having character of interest for each time bin
  pctchint <- rbind(pctchint,getCharPct(chint[[i]], choth[[i]], mstates, msls));
}
#If any values are NaN, turn them to NA so the colMeans will ignore them
#pctchint[is.na(pctchint)] <-  NA;
#Get the average percentages over the set of characters examined
pctchint <- colMeans(pctchint, na.rm=TRUE);
#Plot it
plotCharPct(pctchint,binnames,chintname,chothname,bty="u");
title(main="D",adj=0.05,line=-2, cex.main=1.5, font.main=1);
#Add x axis
axis(side=1, line=1.5,lwd=0.3)
```

```
mtext("Age (Ma)", side=1, line=4.5,cex=par("cex"))
#Close the PDF file with the plots
dev.off();
#Embed fonts
embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font");
#######
```

F.2 R functions called by R script

Code F.2: MorphospaceFunctions.R

```
#Ben Kotrc, Harvard University, December 2011
#kotrc@fas.harvard.edu
#R functions to carry out analysis of diatom morphospace
#These functions are called from the script file "Morphospace.R"

#Function to calculate the completeness of data in a morphospace matrix m
#Takes arguments:
#m - a data matrix, dimensions a*b where a=# of genera and b=# of characters
#inv - string indicating which characters to consider invalid
#      "?" = only count ?
#      "?n" = count ? and n but not v
#      "all" = count ? and n and v
#Returns:
#resDQ - a data frame with components
#resDQ$genus - vector of %ages of valid character states for each genus,
#              names of each element are name of genus
#resDQ$char -  vector of %ages of genera with valid character states for each character,
#              names of each element are name of character (eg. X1, X63, etc.)
getDataQuality <- function(m, inv = "all")
{
  #Calculate, for each character, the %age of taxa that have valid
  #character states, i.e. not ? (state unknown) or n (character
  #not applicable).
  result <- NULL;
  #Count up the number of invalid taxa for each character
  for(i in 1:ncol(m))
  {
    if(inv == "?")
    {
      result[i] <- sum(m[,i] == '?');
    }
    else if(inv == "?n")
    {
      result[i] <- sum(m[,i] == '?') + sum(m[,i] == 'n');
    }
    else if(inv == "NA")
    {
      result[i] <- sum(is.na(m[,i]));
    }
    else
    {
      result[i] <- sum(m[,i] == '?') + sum(m[,i] == 'n') + sum(m[,i] == 'v');
    }
```

Code F.2: MorphospaceFunctions.R (continued)

```
  }
  #Calculate validity %age
  result_percent <- ((nrow(m) - result)/nrow(m)) * 100;
  names(result_percent) <- colnames(m);
  resDQ <- NULL;
  resDQ$char <- result_percent;
  #Calculate, for each genus, the %age of characters that have valid
  #states, i.e. not ? (state unknown) or n (character
  #not applicable).
  result <- NULL;
  #Count up the number of invalid characters for each genus
  for(i in 1:nrow(m))
  {
    if(inv == "?")
    {
      result[i] <- sum(m[i,] == '?');
    }
    else if(inv == "?n")
    {
      result[i] <- sum(m[i,] == '?') + sum(m[i,] == 'n');
    }
    else if(inv == "NA")
    {
      result[i] <- sum(is.na(m[i,]));
    }
    else
    {
      result[i] <- sum(m[i,] == '?') + sum(m[i,] == 'n') + sum(m[i,] == 'v');
    }
  }
  #Calculate validity %age
  result_percent <- ((ncol(m) - result)/ncol(m)) * 100;
  names(result_percent) <- rownames(m);
  resDQ$genus <- result_percent;
  #Send back the results
  return(resDQ);
}


#Function to calculate variances of columns of a matrix
#Taken from https://stat.ethz.ch/pipermail/r-help/2002-March/019606.html
colVars <- function(x, na.rm=FALSE, dims=1, unbiased=TRUE, SumSquares=FALSE,
                    twopass=FALSE) {
  if (SumSquares) return(colSums(x^2, na.rm, dims))
  N <- colSums(!is.na(x), FALSE, dims)
  Nm1 <- if (unbiased) N-1 else N
  if (twopass) {x <- if (dims==length(dim(x))) x - mean(x, na.rm=na.rm) else
    sweep(x, (dims+1):length(dim(x)), colMeans(x,na.rm,dims))}
  (colSums(x^2, na.rm, dims) - colSums(x, na.rm, dims)^2/N) / Nm1
}


#Function to take a morphospace data matrix or data frame
#and replace all the n, v, and ? entries with NA values,
#and return the matrix in "numeric" mode
#Takes m, a matrix or data frame with whatever mode(s)
#Returns m, the matrix in numeric mode with NA values for missing data
makeNumeric <- function(m)
```

Code F.2: MorphospaceFunctions.R (continued)

```
{
  if(mode(m) == "list")
  {
    m <- as.matrix(m);
  }
  #Now we have a matrix of characters, strip out ?, v, and n
  m[m == 'n' | m == 'v' | m == '?'] <- NA;
  #Now switch mode to numeric
  mode(m) <- "numeric";
  return(m);
}


#Function to cull the morphospace data matrix, i.e. take out genera and characters
#that do not meet a threshold % completeness, as calculated by getDataQuality()
#Takes:
#m - the morphospace data matrix to cull
#resDQ - the completeness %ages returned for matrix m by getDataQuality()
#gthresh - the threshold % completeness for genera (e.g. 50)
#cthresh - the threshold % completeness for characters
#Returns:
#mcull - the culled data matrix
cullMatrix <- function(m, resDQ, gthresh=50, cthresh=50)
{
  mcull <- m[names(resDQ$genus[resDQ$genus >= gthresh]),names(resDQ$char[resDQ$char >=
       cthresh])];
  #Send back resulting culled matrix
  return(mcull);
}


#Function to calculate the (g*g) pairwise distance matrix for a given
#(g*c) morphospace data matrix (where g is number of genera, c number of characters)
#Takes:
#m - morphospace data matrix
#Returns:
#d - (pairwise) distance, i.e. dissimilarity, matrix
getDistMatrix <- function(m)
{
  #Store the dissimilarity matrix in "d"
  d <- matrix(0,nrow(m),nrow(m));
  #Go through each taxon, "frow" for focal row
  for (frow in 1:nrow(m))
  {
    #Compare focal row to all other rows, "orow"
    for (orow in 1:nrow(m))
    {
      #Keep track of the number of possible matches (characters
      #for which both taxa have valid states)
      poss_matches <- 0;
      #Keep track of the number of actual matches (characters
      #for which both taxa have the same, valid state)
      matches <- 0;
      #Now go through the characters, i.e. the columns (col)
      for (col in 1:ncol(m))
      {
        #Are both character states valid?
```

Code F.2: MorphospaceFunctions.R (continued)

```
158         if (m[frow,col] != 'n' & m[frow,col] != 'v' & m[frow,col] != '?' & m[orow,col] != '
                n' & m[orow,col] != 'v' & m[orow,col] != '?')
159         {
160           #Increment possible matches
161           poss_matches <- poss_matches + 1;
162           #Are both character states the same?
163           if (m[frow,col] == m[orow,col])
164           {
165             #Increment matches
166             matches <- matches + 1;
167           }
168         }
169       }#End column loop
170       #Now that we have compared each of the characters for
171       #this frow-orow pairing, record the dissimilarity
172       #in the d matrix
173       mismatches <- poss_matches - matches;
174       d[frow,orow] <- mismatches/poss_matches;
175     }#end orow loop
176   }#end frow loop
177   #Add rownames and colnames to matrix
178   rownames(d) <- rownames(m);
179   colnames(d) <- rownames(d);
180   return(d);
181 }
182 
183 
184 #Function to produce a two-dimensional morphospace plot
185 #Takes:
186 #points - x,y pairs to plot
187 #char - a string containing the column label of a character for which to plot non-zero
        states as filled circles
188 #m - a morphospace data matrix (g*c) where g = number of x,y pairs in "points"
189 #sub - a string to be used as the subtitle, none by default
190 #big - whether it's for a main plot or a small plot (Boolean)
191 #namelabs - boolean, determining whether to label points with a string (from rownames of
        points) for their name
192 #Returns:
193 #feedback - a string with error message or positive feedback
194 plotMS2D <- function(points,char=0, m=NULL,sub=NULL,namelabs=FALSE, big=TRUE, axthck=1,
        axcol="grey", bw=FALSE)
195 {
196   #Square plotting region
197   #par(pty="s");
198   #Margins the same width all round
199   if(big == TRUE)
200   {
201     par(mar = c(0, 4, 4, 1));
202     ylab <- "PC 2";
203     xlab <- "PC 1";
204   }
205   else
206   {
207     par(mar = c(0.1,0.1,1,1));
208   }
209   #Generate plot area
210   plot(points,type="n",bty="n", asp=1, main="",col='grey',xaxt="n",yaxt="n",xlab="",ylab=""
        );
```

Code F.2: MorphospaceFunctions.R (continued)

```
  box(which = "plot", lty = "solid", col=axcol, lwd=axthck);
  if(big == TRUE)
  {
    axis(side=3,at=c(-0.1,0,0.1),line=0.5, col=axcol)
    mtext(xlab,side=3,line=3);
    axis(side=2,at=c(-0.1,0,0.1,0.2),line=0.5, col=axcol)
    mtext(ylab,side=2,line=3);
  }
  abline(h=0, lty=1, col=axcol, lwd=axthck);
  abline(v=0, lty=1, col=axcol, lwd=axthck);
  #If there was a request to make fancy plot symbols
  if(char == "Fancy!")
  {
    #Set plot symbol size
    size <- 0.00002;
    for (i in 1:nrow(m))
    {
      #drawRect(points[i,1],points[i,2],size,as.numeric(m[i,"X2"])); #4FCC33
      drawShape(points[i,1],points[i,2],size,as.numeric(m[i,"X1"]),as.numeric(m[i,"X2"]),as
          .numeric(m[i,"X112"]),color="#22990885");
    }
    #Change the subtitle
    #sub <- "Data culled, cmdscale() function, all time bins, symbols refl. aspect ratio &
        shape";
  }
  #Otherwise, if a character of interest was specified
  else if(char != 0)
  {
    #Check to make sure m is the right size for points
    if(dim(points)[1] != dim(m)[1])
    {
      return("m is not the right size for points");
    }
    #Set pch plotting symbols to be used
    symbs <- c(16,22,23,24,24,24);
    #Set colors for plotting symbols
    #cols <- colors()[c(35,563,24,24)];
    cols <- c("#CD3333","#338ACC","#000000","#000000","#000000","#000000");
    if(bw==TRUE) cols <- c("black","black","black","black","black","black");
    bgcols <- colors()[c(35,563,76,24,24,24)];
    if(bw==TRUE) bgcols <- c("black","white","grey","black","black","black")
    #Loop through each valid character state
    for(i in 1:sum(!(levels(as.factor(m[,char])) %in% c("?","n","v"))))
    {
      #This is the current character state
      cur <- levels(as.factor(m[,char]))[!(levels(as.factor(m[,char])) %in% c("?","n","v"))
          ][i];
      #Plot this character state
      points(points[rownames(m[m[,char] == cur,]),1],points[rownames(m[m[,char] == cur,])
          ,2],pch=symbs[i],col=cols[i],cex=0.7,bg=bgcols[i],lwd=0.5);
    }
  }
  #If no character of interest specified
  else
  {
    #Just plot all points
    points(points);
  }
```

Code F.2: MorphospaceFunctions.R (continued)

```
  #Add a subtitle
  mtext(sub, side=3, cex=0.8, padj=-1);
  if(namelabs == TRUE)
  {
    #Add genus names
    text(points,rownames(points),cex=.4,pos=4,offset=.3,col='grey');
  }
  return("Success");
}


#Function to produce a two-dimensional morphospace plot with some genera highlighted
#in specific colors
#Takes:
#points - x,y pairs to plot
#colors - vector of colors with names reflecting genus names of genera to highlight
#namelabs - boolean, determining whether to label points with a string (from rownames of
     points) for their name
#Returns:
#feedback - a string with error message or positive feedback
plotMS2Dgen <- function(points,colors,namelabs=FALSE)
{
  par(mar = c(1,1,1,1));

  #Generate plot area
  plot(points,type="n",bty="n", asp=1, main="",col='grey',xaxt="n",yaxt="n",xlab="",ylab="")
  box(which = "plot", lty = "solid", col="black", lwd=0.3);
  abline(h=0, lty=1, col="black", lwd=0.3);
  abline(v=0, lty=1, col="black", lwd=0.3);

  #Plot nonhighlighted points
  points(points[!(rownames(points) %in% names(colors)),1],points[!(rownames(points) %in%
       names(colors)),2],pch=16,col="grey", cex=1.5);
  #Plot highlighted points
  points(points[(rownames(points) %in% names(colors)),1],points[(rownames(points) %in%
       names(colors)),2],pch=21,col="black",bg=colors,lwd=0.3, cex=1.5);

  if(namelabs == TRUE)
  {
    #Add genus names
    text(points,rownames(points),cex=.4,pos=4,offset=.3,col='grey');
  }
  return("Success");
}


#Function to draw a shape at a given x,y location in an existing plot window
#given some character states
#Takes:
#x,y - location on plot
#size - area (in units squared) of the rectangle surrounding the shape
#ch1, ch2, ch112 - character states (as.numeric) of characters #1, #2 and #112
#color - the color in which  the shape should be plotted
drawShape <- function(x,y,size,ch1,ch2,ch112,color,...)
{
  aspratio <- c(12,4.5,1.75,1.2,1,0.83,0.583,0.25)[ch2+1];
  w <- sqrt(size/aspratio);
```

Code F.2: MorphospaceFunctions.R (continued)

```
  h <- aspratio*w;
  xleft <- x-w;
  xright <- x+w;
  ybottom <- y-h;
  ytop <- y+h;
  if(ch1==1)
  #Rectangle
  {
    rect(xleft,ybottom,xright,ytop,border=NA,col=color,...);
  }
  if(ch1==0)
  #Ellipse
  {
    filledellipse(rx1=w,ry1=h,mid=c(x,y),col=color,lty=0,...);
  }
  if(ch1==2)
  #Rhombus
  {
    filledmultigonal(nr=4,rx=w,ry=h,mid=c(x,y),col=color,lty=0,...);
  }
  if(ch1==3)
    #Oval (ish)
  {
    #Latissime through perdepresse (then rotate °90)?
    if(ch2 %in% c(5,6,7))
    {
      #Left half
      filledellipse(rx1=h/2,ry1=w,mid=c(x,y),col=color,lty=0,from=pi/2,to=(3/2)*pi,...);
      #Right half
      filledellipse(rx1=(3/2)*h,ry1=w,mid=c(x,y),col=color,lty=0,from=(3/2)*pi,to=pi/2,...)
          ;
      #Pass width down as length for raphe plotting
      h <- w;
    }
    #Oval (not rotated)
    else
    {
      #Bottom half
      filledellipse(rx1=w,ry1=h/2,mid=c(x,y-(h/2)),col=color,lty=0,from=pi,to=2*pi,...);
      #Top half
      filledellipse(rx1=w,ry1=(3/2)*h,mid=c(x,y-(h/2)),col=color,lty=0,from=0,to=pi,...);
    }

  }
  if(ch1==4)
  #Triangle
  {
    filledmultigonal(nr=3,rx=w*1.2,ry=h*1.2,mid=c(x,y),angle=210,col=color,lty=0,...);
  }
  #Raphe?
  if(ch112 %in% c(1,2,3,4))
  {
    if(color==”#22990885”)
    {
      segments(x,y-(h/2),x,y+(h/2),col=”darkgreen”,...);
    } else
    {
      segments(x,y-(h/2),x,y+(h/2),col=”black”,...);
```

Code F.2: MorphospaceFunctions.R (continued)

```
    }
  }
}


#Function to produce a two-dimensional morphospace plot annotated with images around the margins
#Takes:
#points - list of x,y coordinates
#images - string vector containing the list of names to plot
plotMS2DImages <- function(points, name, savePDF=FALSE, axthck=1)
{
  #Plot area limits
  xrange <- c(-0.3,0.3);
  yrange <- c(-0.3,0.3);
  #Generate basic plot
  par(mar=c(0, 4, 5, 0) + 0.2, pty="s");
  plot(pco$points,type="p", bty="n",xaxs="i",yaxs="i",xaxt="n",yaxt="n",xlab="", ylab="", pch=16, col='black', main="",xlim=xrange,ylim=yrange);
  axis(side=3,at=c(-0.1,0,0.1),line=1.5, col="grey",lwd=axthck);
  mtext("PC 1",side=3,line=4);
  axis(side=2,at=c(-0.1,0,0.1,0.2),line=0.5, col="grey", lwd=axthck);
  mtext("PC 2",side=2,line=3);
  segments(0,-0.1660883, 0,0.2335540, lty=1, col="grey",lwd=axthck);
  segments(-0.175, 0, 0.1759463,0, lty=1, col="grey",lwd=axthck);
  rect(-0.175, -0.1660883,  0.1759463,  0.2335540, border="grey",lwd=axthck);
  points(pco$points, pch=16, col='black');
  #par(col="black");
  #How many pictures along the left and right, total?
  sides <- 4;
  #Width of each image in plot units
  width <- diff(xrange)/(((length(name)-sides)/2)+1);
  #Width between images
  wspacing <- diff(xrange)/((length(name)-sides)/2);
  #Height between images
  hspacing <- diff(yrange)/((sides/2)+2);
  #Plot the top row
  for (i in 0:ceiling(((length(name)-sides)/2)-1))
  {
    #Load matching image
    image <- read.jpeg(paste("images/", name[i+1], ".jpg", sep=""));
    #Image aspect ratio
    aspect <- dim(image)[1]/dim(image)[2];
    #Set image corner coordinates
    xleft <- xrange[1]+(i*wspacing);
    ytop <- yrange[2]+0.03;
    ybottom <- ytop-(width*aspect);
    xright <- xleft+width;
    #Plot 'er up!
    rasterImage(image,xleft,ybottom,xright,ytop,xpd=TRUE);
    #Connect with a line
    segments((xleft+(width/2)),ybottom,pco$points[row.names(pco$points)==name[i+1]][1],pco$points[row.names(pco$points)==name[i+1]][2],lwd=axthck);
    #text(xleft,ybottom,name[i+1]);
    }
  #Now plot the right side
  for (i in 1:(sides/2))
  {
```

Code F.2: MorphospaceFunctions.R (continued)

```
    #Load matching image
    image <- read.jpeg(paste("images/", name[((length(name)-sides)/2)+i], ".jpg", sep=""));
    #Image aspect ratio
    aspect <- dim(image)[1]/dim(image)[2];
    #Set image corner coordinates
    xleft <- xrange[2]-wspacing;
    xright <- xleft+width;
    ytop <- yrange[2]-(i)*hspacing;
    ybottom <- ytop-(width*aspect);
    #Plot 'er up!
    rasterImage(image,xleft,ybottom,xright,ytop);
    #Connect with a line
    segments(xleft,(ybottom+(width*aspect)/2),pco$points[row.names(pco$points)==name[((
        length(name)-sides)/2)+i]][1],pco$points[row.names(pco$points)==name[((length(name
        )-sides)/2)+i]][2],lwd=axthck);
    #text(xleft,ybottom,name[((length(name)-sides)/2)+i]);
    }
  #Plot the bottom row
  for (i in 0:ceiling(((length(name)-sides)/2)-1))
  {
    #Load matching image
    image <- read.jpeg(paste("images/", name[ceiling(length(name)-(sides/2))-i], ".jpg",
        sep=""));
    #Image aspect ratio
    aspect <- dim(image)[1]/dim(image)[2];
    #Set image corner coordinates
    xleft <- xrange[1]+(i*wspacing);
    xright <- xleft+width;
    ybottom <- yrange[1]+0.05;
    ytop <- ybottom+(width*aspect);
    #Plot 'er up!
    rasterImage(image,xleft,ybottom,xright,ytop);
    #Connect with a line
    segments((xleft+(width/2)),ytop,pco$points[row.names(pco$points)==name[ceiling(length(
        name)-(sides/2))-i]][1],pco$points[row.names(pco$points)==name[ceiling(length(name
        )-(sides/2))-i]][2],lwd=axthck);
    #text(xleft,ybottom,name[ceiling(length(name)-(sides/2))-i]);
    }
  #Finally, plot the left side
  for (i in 1:(sides/2))
  {
    #Load matching image
    image <- read.jpeg(paste("images/", name[(length(name)-(i-1))], ".jpg", sep=""));
    #Image aspect ratio
    aspect <- dim(image)[1]/dim(image)[2];
    #Set image corner coordinates
    xleft <- xrange[1];
    xright <- xleft+width;
    ytop <- yrange[2]-(i)*hspacing;
    ybottom <- ytop-(width*aspect);
    #Plot 'er up!
    rasterImage(image,xleft,ybottom,xright,ytop);
    #Connect with a line
    segments(xright,(ybottom+(width*aspect)/2),pco$points[row.names(pco$points)==name[(
        length(name)-(i-1))]][1],pco$points[row.names(pco$points)==name[(length(name)-(i
        -1))]][2],lwd=axthck);
    #text(xleft,ybottom,name[(length(name)-(i-1))]);
    }
```

Code F.2: MorphospaceFunctions.R (continued)

```
  #Save a PDF version?
  if (savePDF == TRUE)
  {
    #Get the fonts ready
    file.exists <- function( fname ) length(Sys.glob(fname))>0
    absolute.path.to.font.files <-  "/Users/bkotrc/font/";
    bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
    ## if you do not have the correct font types
    for (i in 1:length(bera.names)) {
      stopifnot( file.exists(paste(absolute.path.to.font.files,
                                   bera.names[i], ".afm", sep="")) )
      stopifnot( file.exists(paste(absolute.path.to.font.files,
                                   bera.names[i], ".otf", sep="")) )
    }
    gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files,bera.names, ".afm",
          sep=""))
    pdfname <- "pco_images.pdf";
    #Make a composite plot of previous plus this plot for publication
    pdf(file=pdfname, bg="white", width=(7.2/cm(1)), height=(7.2/cm(1)), pointsize=7,
        family=gillsans);
    par(font.main=1, cex.main=1.5)
    axthck <- 0.3;
    #Make a recursive call to this function, without the savePDF parameter, which defaults
        to FALSE
    plotMS2DImages(points=points,name=name, axthck=axthck);
    #Close PDF file
    dev.off();
    #Now embed font in that file
    embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font");
  }
}


#Function to prepare Neptune database for use with morphospace by creating a new column
    with the
#diatom Genus name only and fixing mistyped names
#Takes:
#N - a Neptune database with the usual columns
#Returns:
#N - a Neptune database with an extra $Genus column with only the genus name, and some
    names changed
prepNeptune <- function(N)
{
  #Now, let's split out the genus names on their own
  #(No need to do this each –timein future load modified file at top)
  for (i in 1:dim(N)[1]){
          N$Genus[i] <- strsplit(as.character(N$Species[i])," ")[[1]][1];
  }
  #Now save to file for subsequent use
  write.table(N, file='NeptuneGenNames.txt', sep="\t");
  #We need to fix some Genus names that are incorrect in Neptune
  #Read in the file, but don't convert strings to factors
  N <- read.table(file='NeptuneGenNames.txt', header=TRUE, sep="\t",as.is=TRUE);
  #Print a list of the Genus names
  #sort(unique(N$Genus));
  #Now change a whole bunch of names
  N[N$Genus=='Aulacosira',]$Genus <- 'Aulacoseira'
  N[N$Genus=='Lisitzina',]$Genus <- 'Lisitzinia'
```

Code F.2: MorphospaceFunctions.R (continued)

```
  N[N$Genus==’Bacteriosira’,]$Genus <- ’Bacterosira’
  N[N$Genus==’Bruniopsis’,]$Genus <- ’Neobrunia’
  N[N$Genus==’Calloneis’,]$Genus <- ’Caloneis’
  N[N$Genus==’Denticulopsos’,]$Genus <- ’Denticulopsis’
  N[N$Genus==’Neohuttonia’,]$Genus <- ’Huttonia’
  N[N$Genus==’Neodelphines’,]$Genus <- ’Neodelphineis’
  N[N$Genus==’Opephoneis’,]$Genus <- ’Opephora’
  N[N$Genus==’Pseudostitodiscus’,]$Genus <- ’Pseudostictodiscus’
  N[N$Genus==’Raphidodiscus’,]$Genus <- ’Rhaphidodiscus’
  N[N$Genus==’Screptroneis’,]$Genus <- ’Sceptroneis’
  N[N$Genus==’Simonsenella’,]$Genus <- ’Simonseniella’
  N[N$Genus==’Stephanophyxis’,]$Genus <- ’Stephanopyxis’
  N[N$Genus==’Stephonopyxis’,]$Genus <- ’Stephanopyxis’
  N[N$Genus==’Stichodiscus’,]$Genus <- ’Stictodiscus’
  N[N$Genus==’Thalassoithrix’,]$Genus <- ’Thalassiothrix’
  N[N$Genus==’Dactyliozolen’,]$Genus <- ’Dactyliosolen’
  N[N$Genus==’Cosinodiscus’,]$Genus <- ’Coscinodiscus’
  N[N$Genus==’Pinnularia’,]$Genus <- ’Caloneis’
  N[N$Genus==’Rhabdonoma’,]$Genus <- ’Rhabdonema’
  #Dicladia is a resting cell
  #Pseudopyxilla is a resting cell
  #Muelleriopsis is a resting cell
  #Liradiscus is a resting cell
  #Odontotropis is a resting cell
  #Chasea is a resting cell
  N[N$Genus==’Coscinodixcus’,]$Genus <- ’Coscinodiscus’
  N[N$Genus==’Arachnodiscus’,]$Genus <- ’Arachnoidiscus’
  N[N$Genus==’Pterothecas’,]$Genus <- ’Pterotheca’
  return(N);
}


#Function to make a list of taxa in time bins
#Takes:
#N - a Neptune database
#bins - a vector containing the boundaries between adjacent time bins
#mode - ”range-through” or ”in-bin”
#(not including the beginning and end of the dataset)
#Retuns:
#msls - a data frame of Genus lists, one for each time bin (i.e. length(bins)+1)
getBinLists <- function(N, bins, binnames, mode=”range-through”)
{
  if(mode==”in-bin”)
  {
    #In-bin sampling:
    #Data fram to contain lists for each time bin
    msls <- NULL;
    msls <- as.list(msls);
    #First (oldest) time bin
    if(length(unique(N[N$Sample.Age > bins[1],]$Genus)) == 0)
    {
      msls[[1]] <- NA
    }else
    {
      msls[[1]] <- as.character(unique(N[N$Sample.Age > bins[1],]$Genus));
    }
    #Loop through middle time bins
    for(i in 1:(length(bins)-1))
```

Code F.2: MorphospaceFunctions.R (continued)

```
    {
      #What if the time bin is empty?
      if(length(unique(N[(N$Sample.Age <= bins[i] & N$Sample.Age > bins[i+1]),]$Genus)) ==
          0)
      {
         msls[[i+1]] <- NA;
      }else
      {
        msls[[i+1]] <- as.character(unique(N[(N$Sample.Age <= bins[i] & N$Sample.Age > bins
            [i+1]),]$Genus));
      }
    }
    #Last (youngest) time bin
    msls[[(length(bins)+1)]] <- as.character(unique(N[N$Sample.Age <= bins[length(bins)],]$
        Genus));
  }
  else if(mode==”range-through”)
  {
    #Range-through sampling:
    #Data fram to contain lists for each time bin
    msls <- NULL;
    as.list(msls);
    #First (oldest) time bin
    msls[[1]] <- as.character(unique(N[N$Sample.Age > bins[1],]$Genus));
    #Loop through middle time bins
    for(i in 1:(length(bins)-1))
    {
      #Define start (max. age) and end (min. age) of time bin
      binstart <- bins[i];
      binend <- bins[i+1];
      #Divide Neptune into before, during, and after time bin
      beforebin <- N[N$Sample.Age > binstart,];
      focalbin <- N[N$Sample.Age <= binstart & N$Sample.Age > binend,];
      afterbin <- N[N$Sample.Age <= binend,];
      #Get list of taxa in focal & pre/post bins
      focalset <- unique(focalbin$Genus);
      beforeset <- unique(beforebin$Genus);
      afterset <- unique(afterbin$Genus);
      #Range-through list is combination (union) of focal bin with those taxa in both
      #the before and after bin (i.e. the interesection of those bins)
      list <- union(intersect(beforeset,afterset),focalset);
      #Also store for later stacked plotting
      msls[[i+1]] <- list;
    }
    #Last (youngest) time bin
    msls[[(length(bins)+1)]] <- as.character(unique(N[N$Sample.Age <= bins[length(bins)],]$
        Genus));
  }
  #Label the resulting lists in msls with the bin names
  names(msls) <- binnames;
  return(msls);
}


#Function to generate a 3D morphospace plot
#Takes:
#points - x,y locations of points to be plotted, with rows labeled as genus names
#msls - list of lists genus names in each time bin
```

Code F.2: MorphospaceFunctions.R (continued)

```
#savePDF - boolean, whether to output to PDF file in current wd
#mode - "color by bin" or "linking"
#m - a morphospace matrix, if mode is "linking"
plotMS3D <- function(points, msls, savePDF=FALSE, mode="color by bin", m=NULL, sizes=1)
{
  #Set color scheme for time slices
  ecretc <- rgb(red=140,green=205,blue=87,alpha=200,maxColorValue=255);
  lcretc <- rgb(red=166,green=216,blue=74,alpha=200,maxColorValue=255);
  pacc <- rgb(red=253,green=167,blue=95,alpha=200,maxColorValue=255);
  eocc <- rgb(red=253,green=180,blue=108,alpha=200,maxColorValue=255);
  olic <- rgb(red=253,green=192,blue=122,alpha=200,maxColorValue=255);
  mioc <- rgb(red=255,green=255,blue=0,alpha=200,maxColorValue=255);
  plioc <- rgb(red=255,green=255,blue=153,alpha=200,maxColorValue=255);
  pleic <- rgb(red=255,green=242,blue=174,alpha=200,maxColorValue=255);
  #If Cenozoic only:
  if(names(msls)[1]=="Paleocene")
  {
    polycols <- c(pacc,eocc,olic,mioc,plioc,pleic);
  }else
  {
  polycols <- c(ecretc,lcretc,pacc,eocc,olic,mioc,plioc);
  }
  #Get list of all taxa
  all <- NULL;
  for(i in 1:length(msls))
  {
    all <- c(all,msls[[i]]);
  }
  all <- unique(all);
  #Pseudorutilaria is is way off the scale, so let's set up the plot area without it.
  almostall <- all[all != "Pseudorutilaria"];
  #Now set up the axes
  par(lwd=0.3)
  #If only plotting Cenozoic, make bottom margin wider to keep the time
  #bin labels close to the slices for visual niceness
if(names(msls)[1]=="Paleocene")
  {
    #Only Cenozoic slices plotted
  ms3d <- scatterplot3d(cbind(points[almostall,],0),type="n",color="black",pch=16,zlim=c(0,
       length(msls)),xlim=range(points[almostall,1]),ylim=range(points[almostall,2]),angle
       =48,xlab="PC1",ylab="",zlab="Time Bin",main="",box=FALSE,z.ticklabs=c(rep("",length(
       msls))),x.ticklabs=c("-.2","0",".2"),y.ticklabs=c("-.2","0",".2"),lab=c(2,2,1),lab.z
       =length(msls),xpd=TRUE,mar=c(6.5,2.5,0,0)+0.1);
  }else
  {
    #Plotting Mesozoic and Cenozoic slices
    ms3d <- scatterplot3d(cbind(points[almostall,],0),type="n",color="black",pch=16,zlim=c
         (0,length(msls)),xlim=range(points[almostall,1]),ylim=range(points[almostall,2]),
         angle=48,xlab="PC1",ylab="",zlab="Time Bin",main="",box=FALSE,z.ticklabs=c(rep("",
         length(msls))),x.ticklabs=c("-.2","0",".2"),y.ticklabs=c("-.2","0",".2"),lab=c
         (2,2,1),lab.z=length(msls),xpd=TRUE,mar=c(2.5,2.5,0,0)+0.1);
  }
  #Add translucent polygons showing occupied morphospace at z=0
  #First, calculate convex hulls surrounding each set of points
  polys <- getHulls(points,msls);
  #Now convert points so they'll plot correctly in 3D, then plot
  epochs <- names(msls);
  #Loop through time bins, plot z=0 hulls
```

Code F.2: MorphospaceFunctions.R (continued)

```
  for (i in 0:(length(msls)-1))
  {
    v<-ms3d$xyz.convert(polys[[length(epochs)-i]][,1], polys[[length(epochs)-i]][,2], rep
        (0,length(polys[[length(epochs)-i]][,1])));
    polygon(v,col=polycols[length(epochs)-i],lwd=1);
  }
  #Now add planes for each time slice except the bottom
  for (i in 1:length(epochs))
  {
    #Solid, mostly opaque white plane
    v<-ms3d$xyz.convert(c(-0.2,0.2,0.2,-0.2), c(-0.2,-0.2,0.2,0.2), rep(i, times=4));
    polygon(v,col="#FFFFFF95",lty="blank");
    #Outline of plane
    ms3d$plane3d(i,0,0, lty="solid", lwd=0.3, col="black");
    #ms3d$points3d(0,0,i,pch=16,col="#00000000",cex=1);
  }

  #Label the time slices
  coords <- c(0.185,0.232,1);
  #Shorten Cretaceous binnames
  #epochs[1:2] <- c("E. Cretaceous", "L. Cretaceous");
  for(i in 2:length(epochs))
  {
    coords <- rbind(coords,c(0.185,0.232,i));
  }
  placement <- ms3d$xyz.convert(coords);
  #Set rectangle expansion
  exp <- 1.5;
  #Add colored rectangles
  rect(placement$x-(strwidth(epochs)),placement$y-((strheight(epochs)*exp)/2),placement$x,
       placement$y+((strheight(epochs)*exp)/2),col=polycols,border=NA);
  #Add text labels
  text(placement,labels=epochs,adj=1);

  placement2 <- ms3d$xyz.convert(0.35,-0.15,0);
  text(placement2, "PC2");
  #If plotting sized by occurrence, provide key to each time slice's size
  if(length(sizes)!=1)
  {
    points(placement$x-0.03,placement$y-0.4,col="black",bg=polycols,pch=21,lwd=0.3)
    text(placement$x-0.07,placement$y-0.4,labels=round(sizes$meanoccs),adj=1,cex=0.7);
  }
  if(mode == "color by bin")
  {
    #Now add points for each time slice, color coded by time bin
    #Loop through each time bin
    for (i in 1:length(epochs))
    {
      #if not specifying plot point sizes
      if(length(sizes)==1)
      {
        ms3d$points3d(cbind(points[msls[[i]],],i),pch=16,col=polycols[i],cex=1);
      ms3d$points3d(cbind(points[msls[[i]],],i),pch=1,col="black",lwd=0.3,cex=1);

      }else
      {
        ms3d$points3d(cbind(points[msls[[i]],],i),pch=16,col=polycols[i],cex=sizes[[i]]);
      ms3d$points3d(cbind(points[msls[[i]],],i),pch=1,col="black",lwd=0.3,cex=sizes[[i]]);
```

Code F.2: MorphospaceFunctions.R (continued)

```
753
754       }
755           }
756   }
757   else if(mode == "linking")
758   {
759     #Loop through each time bine
760     for (i in 1:length(epochs))
761     {
762       #Genera have linking if non-zero states for characters 28, 54, 55, 59, or 96
763       res28 <- hasChar(msls[[i]], "X28", m);
764       res54 <- hasChar(msls[[i]], "X54", m);
765       res55 <- hasChar(msls[[i]], "X55", m);
766       res59 <- hasChar(msls[[i]], "X59", m);
767       res96 <- hasChar(msls[[i]], "X96", m);
768       nonzero <- unique(c(res28$nonzero,res54$nonzero,res55$nonzero,res59$nonzero,res96$
              nonzero));
769       zero <- setdiff(unique(c(res28$zero,res54$zero,res55$zero,res59$zero,res96$zero)),
              nonzero);
770       #Plot nonzero states as black circles
771       ms3d$points3d(cbind(points[nonzero,],i),pch=16,col="black");
772       #Plot zero states as hollow circles
773       ms3d$points3d(cbind(points[zero,],i),pch=1,col="black",lwd=0.5);
774     }
775   }
776   #Save PDF?
777   if(savePDF == TRUE)
778   {
779     #Get the fonts ready
780   file.exists <- function( fname ) length(Sys.glob(fname))>0
781   absolute.path.to.font.files <-  "/Users/bkotrc/font/";
782   bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
783   ## if you do not have the correct font types
784   for (i in 1:length(bera.names)) {
785     stopifnot( file.exists(paste(absolute.path.to.font.files,bera.names[i], ".afm", sep="")
            ) )
786     stopifnot( file.exists(paste(absolute.path.to.font.files,bera.names[i], ".otf", sep="")
            ) )
787   }
788   gillsans <- Type1Font("GillSans",paste(absolute.path.to.font.files,bera.names, ".afm",
          sep=""))
789   pdfname <- "Untitled_3D_plot.pdf";
790   #Plot it
791   pdf(file=pdfname, bg="white", width=(7.2/cm(1)), height=(14.4/cm(1)), pointsize=7, family
          =gillsans, colormodel="cmyk");
792   plotMS3D(points,msls,mode=mode,m=m,sizes=sizes);
793     dev.off();
794     #Now embed font in that file
795   embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font");
796   }
797 }
798
799
800 #Function to generate convex hulls for sets of morphospace points in a plot
801 #Takes:
802 #points - x,y locations of points to be plotted, with rows labeled as genus names
803 #msls - list of lists genus names in each time bin
804 #Returns:
```

Code F.2: MorphospaceFunctions.R (continued)

```
#polys - a list of x,y locations of convex hulls surrounding the genera in each time bin
getHulls <- function(points, msls)
{
  polys <- NULL;
  polys <- as.list(polys);
  for (i in 1:length(msls))
  {
    #Calculate the polygon enclosing those points
    hull <- chull(points[msls[[i]],]);
    polys[[i]]  <- points[msls[[i]],][hull,];
  }
  return(polys);
}


#Function to determine whether the genera in time bin lists "msls" contain
#valid non-zero states for the characters in "char"
#Takes
#msls - a single list (vector of strings) of genus names
#char - a morphospace character of interest (e.g. "X60")
#m - a data matrix, dimensions a*b where a=# of genera and b=# of characters
#Returns components of res, res$...
#invalid - list of genus names for which the character state is n, v, or ?
#zero - list of genus names for which the character state is 0
#nonzero - list of genus names for which the character state is valid but not 0
hasChar <- function(msls, char, m)
{
  #Result will be in data frame
  res <- NULL;
  #Invalid characters
  invset <- c("n","v","?");
  #Set of character state for char that are not 0 or invalid
  nzset <- setdiff(unique(m[,char]),c("0",invset));
  #List of genera with state 0 for char
  res$zero <- intersect(rownames(m[m[,char] == 0,]),msls);
  #List of genera with state n, v, or ? for char
  res$invalid <- intersect(rownames(m[m[,char] %in% invset,]),msls);
  #The rest
  res$nonzero <- intersect(rownames(m[m[,char] %in% nzset,]),msls);

  return(res);
}


#Function to calculate the mean pairwise distance for a list of taxa
#Takes
#d - a distance matrix with named rows and columns
#msls - a list of lists of names, also found in d, for which to calculate the mean pairwise
      distance
#Returns
#result - vector of mean pairwise distances for the time bin lists submitted to function
meanPairwiseDist <- function(d, msls)
{
  #Set up results vector
  result <- vector(mode="numeric",length=length(msls));
  names(result) <- names(msls);

  #Loop over each time bin
```

Code F.2: MorphospaceFunctions.R (continued)

```
  for(i in 1:length(msls))
  {
    if((is.na(msls[[i]])) != TRUE)
    {
    #Extract relevant part of d matrix
    di <- d[msls[[i]],msls[[i]]];
    #Get distance matrix as lower triangular
    dlower <- di[lower.tri(di)];
    #Now get mean of all values in lower triangular
    result[i] <- mean(dlower);
    }
    else
    {
      result[i] <- NA;
    }
  }
  return(result);
}


#Function to calculate the convex hull volume for a list of lists of taxa
#Takes
#p - matrix with m rows of genera describing their location in n-dimensional PCO space
#msls - a list of lists of names, also found in d, for which to calculate the mean pairwise
      distance
#dim - how many of the n dimensions to consider
#Returns
#result - vector of convex hulls for the time bin lists submitted to function
convexHullVol <- function(p, msls, dim)
{
  #Set up results vector
  result <- vector(mode="numeric",length=length(msls));
  names(result) <- names(msls);

  #Loop over each time bin
  for(i in 1:length(msls))
  {
    if((is.na(msls[[i]])) != TRUE)
    {
    #Extract relevant part of p, the PCO points matrix
    pi <- p[msls[[i]],1:dim];
    #Get volume of convex hull around points in pi
    result[i] <- convhulln(pi, options="FA")[[3]];
    }else
    {
      result[i] <- NA;
    }
  }
  return(result);
}


#Function to calculate the alpha shape volume for a list of lists of taxa
#Takes
#points - matrix with m rows of genera describing their location in 3-dimensional PCO space
#msls - a list of lists of names, also found in d, for which to calculate the mean pairwise
      distance
#alpha - alpha value to use in constructing alpha shape
```

Code F.2: MorphospaceFunctions.R (continued)

```
#Returns
#result - vector of alpha shape volumes for the time bin lists submitted to function
alphaShapeVol <- function(alpha,points,msls)
{
  #Set up results vector
  result <- vector(mode="numeric",length=length(msls));
  names(result) <- names(msls);

  #Loop over each time bin
  for(i in 1:length(msls))
  {
    if((is.na(msls[[i]])) != TRUE)
    {
    #Extract relevant part of p, the PCO points matrix
    pi <- points[msls[[i]],1:3];
    #Get the alpha shape for the selected point cloud as an "ashape3d" class object
    alphaobj <- ashape3d(pi,alpha=alpha);
    #Calculate volume of the alpha shape
    result[i] <- volume_ashape3d(alphaobj);
    }
    else
    {
      result[i] <- NA;
    }
  }
  return(result);
}


#Function to figure with seven character state distributions in PCO1-2
plotSevenPartFig <- function(pco,mfull,bw=FALSE)
{
  #Get the fonts ready
  file.exists <- function( fname ) length(Sys.glob(fname))>0
  absolute.path.to.font.files <-  "/Users/bkotrc/font/";
  bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
  ## if you do not have the correct font types
  for (i in 1:length(bera.names)) {
    stopifnot( file.exists(paste(absolute.path.to.font.files,
                                 bera.names[i], ".afm", sep="")) )
    stopifnot( file.exists(paste(absolute.path.to.font.files,
                                 bera.names[i], ".otf", sep="")) )
  }
  gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files,
                                          bera.names, ".afm", sep=""))
  pdfname <- "indivcharpco.pdf";
  #Make a composite plot of previous plus this plot for publication
  pdf(file=pdfname, bg="white", width=(7.2/cm(1)), height=(21/cm(1)), pointsize=7, family=
      gillsans);
  #Text settings
  abc <- gpar(fontface=1, cex=1.2);
  bold <- gpar(fontface=1, cex=1);
  plain <- gpar(fontface=3, cex=0.8);
  red <- gpar(col="#CD3333",fill="#CD3333",cex=0.4);
  blue <- gpar(col="#338ACC",fill="#338ACC",cex=0.4);
  yellow <- gpar(col="black",cex=0.4,fill="#FFB90F",lwd=0.5);
  black <- gpar(col="black",cex=0.4,fill="black",lwd=0.5);
  grey <- gpar(col="grey",cex=0.8);
```

Code F.2: MorphospaceFunctions.R (continued)

```
  if(bw==TRUE)
  {
    red <- gpar(col="black",cex=0.4);
    blue <- gpar(col="black",fill="white",cex=0.4,lwd=0.5);
    yellow <- gpar(col="black",fill="grey",cex=0.4,lwd=0.5);
  }
  axthck=0.3;
  axcol="black";
  #Start with character X60 (pennate vs. centric)
  par(fig=c(2/72,32/72,6/7,1));
  plotMS2D(pco$points[,1:2],char="X60",m=mfull[rownames(d),],namelabs=FALSE,big=FALSE,
       axthck=axthck,axcol=axcol,bw=bw);
  #Add graphic legend (using grid)
  vp <- viewport(x=46/72, y=6.5/7, width=30/72, height=30/210, angle=0);
  pushViewport(vp);
  r <- readPNG("fig1labels/x60.png");
  grid.raster(r);
  grid.text("Shape of structural pattern center (48)",x=0,y=0.9, just="left", gp=bold);
  grid.text("Ring-shaped (annulus)",x=0.05,y=0.77, just="left", gp=plain);
  grid.points(x=0.02,y=0.77,pch=16,gp=red);
  grid.text("Linear (sternum)",x=0.05,y=0.3, just="left", gp=plain);
  grid.points(x=0.02,y=0.3,pch=22,gp=blue);
  popViewport();
  grid.text("A", x=4/72, y=204.5/210, gp=abc, just="left");
  grid.text("PC 1", x=4/72, y=208.5/210, gp=grey, just="left");
  grid.text("0", x=15/72, y=208.5/210, gp=grey, just="left");
  grid.text("PC 2", x=1/72, y=201/210, gp=grey, just="left", rot=90);
  grid.text("0", x=0.5/72, y=191.5/210, gp=grey, just="left");
  #Next, character X34
  par(fig=c(2/72,32/72,5/7,6/7),new=TRUE);
  plotMS2D(pco$points[,1:2],char="X34",m=mfull[rownames(d),],namelabs=FALSE,big=FALSE,
       axthck=axthck,axcol=axcol,bw=bw);
  vp <- viewport(x=46/72, y=5.5/7, width=30/72, height=30/210, angle=0);
  pushViewport(vp);
  r <- readPNG("fig1labels/x34.png");
  grid.raster(r);
  grid.text("Mantle curvature (27)",x=0,y=0.9, just="left", gp=bold);
  grid.text("Straight",x=0.05,y=0.77, just="left", gp=plain);
  grid.points(x=0.02,y=0.77,pch=16,gp=red);
  grid.text("Convex",x=0.05,y=0.38, just="left", gp=plain);
  grid.points(x=0.02,y=0.38,pch=22,gp=blue);
  popViewport();
  grid.text("B", x=4/72, y=(204.5-30)/210, gp=abc, just="left");
  #Now, character X26
  par(fig=c(2/72,32/72,4/7,5/7),new=TRUE);
  plotMS2D(pco$points[,1:2],char="X26",m=mfull[rownames(d),],namelabs=FALSE,big=FALSE,
       axthck=axthck,axcol=axcol,bw=bw);
  vp <- viewport(x=46/72, y=4.5/7, width=30/72, height=30/210, angle=0,);
  pushViewport(vp);
  r <- readPNG("fig1labels/x26.png");
  grid.raster(r);
  grid.text("Angle betw. valve face and mantle (19)",x=0,y=0.9, just="left", gp=bold);
  grid.text("No clear distinction",x=0.05,y=0.77, just="left", gp=plain);
  grid.points(x=0.02,y=0.77,pch=16,gp=red);
  grid.text("Right angle",x=0.05,y=0.49, just="left", gp=plain);
  grid.points(x=0.02,y=0.49,pch=22,gp=blue);
  grid.text("Obtuse angle",x=0.05,y=0.2, just="left", gp=plain);
  grid.points(x=0.02,y=0.2,pch=23,gp=yellow);
```

Code F.2: MorphospaceFunctions.R (continued)

```
1030  popViewport();
1031  grid.text("C", x=4/72, y=(204.5-60)/210, gp=abc, just="left");
1032  #Then character X67
1033  par(fig=c(2/72,32/72,3/7,4/7),new=TRUE);
1034  plotMS2D(pco$points[,1:2],char="X67",m=mfull[rownames(d),],namelabs=FALSE,big=FALSE,
          axthck=axthck,axcol=axcol,bw=bw);
1035  vp <- viewport(x=46/72, y=3.5/7, width=30/72, height=30/210, angle=0,);
1036  pushViewport(vp);
1037  r <- readPNG("fig1labels/x67.png");
1038  grid.raster(r);
1039  grid.text("Uniformity of pore size (54)",x=0,y=0.9, just="left", gp=bold);
1040  grid.text("Uniform",x=0.05,y=0.77, just="left", gp=plain);
1041  grid.points(x=0.02,y=0.77,pch=16,gp=red);
1042  grid.text("Larger at",x=0.55,y=0.77, just="left", gp=plain);
1043  grid.text("pattern center",x=0.55,y=0.70, just="left", gp=plain);
1044  grid.points(x=0.52,y=0.77,pch=22,gp=blue);
1045  grid.text("Smaller at",x=0.05,y=0.35, just="left", gp=plain);
1046  grid.text("pattern center",x=0.05,y=0.29, just="left", gp=plain);
1047  grid.points(x=0.02,y=0.35,pch=23,gp=yellow);
1048  grid.text("Irregular",x=0.55,y=0.35, just="left", gp=plain);
1049  grid.points(x=0.52,y=0.35,pch=24,gp=black);
1050  popViewport();
1051  grid.text("D", x=4/72, y=(204.5-90)/210, gp=abc, just="left");
1052  #Character X27
1053  par(fig=c(2/72,32/72,2/7,3/7),new=TRUE);
1054  plotMS2D(pco$points[,1:2],char="X27",m=mfull[rownames(d),],namelabs=FALSE,big=FALSE,
          axthck=axthck,axcol=axcol,bw=bw);
1055  vp <- viewport(x=46/72, y=2.5/7, width=30/72, height=30/210, angle=0,);
1056  pushViewport(vp);
1057  r <- readPNG("fig1labels/x27.png");
1058  grid.raster(r);
1059  grid.text("Ornament at rim (20)",x=0,y=0.9, just="left", gp=bold);
1060  grid.text("None",x=0.1,y=0.82, just="left", gp=plain);
1061  grid.points(x=0.07,y=0.82,pch=16,gp=red);
1062  grid.text("Simple short",x=0.55,y=0.82, just="left", gp=plain);
1063  grid.text("spinules",x=0.55,y=0.76, just="left", gp=plain);
1064  grid.points(x=0.52,y=0.82,pch=22,gp=blue);
1065  grid.text("Long spines or",x=0.1,y=0.4, just="left", gp=plain);
1066  grid.text("marginal processes",x=0.1,y=0.33, just="left", gp=plain);
1067  grid.points(x=0.07,y=0.4,pch=23,gp=yellow);
1068  popViewport();
1069  grid.text("E", x=4/72, y=(204.5-120)/210, gp=abc, just="left");
1070  #Character X61
1071  par(fig=c(2/72,32/72,1/7,2/7),new=TRUE);
1072  plotMS2D(pco$points[,1:2],char="X61",m=mfull[rownames(d),],namelabs=FALSE,big=FALSE,
          axthck=axthck,axcol=axcol,bw=bw);
1073  vp <- viewport(x=46/72, y=1.5/7, width=30/72, height=30/210, angle=0,);
1074  pushViewport(vp);
1075  r <- readPNG("fig1labels/x61.png");
1076  grid.raster(r);
1077  grid.text("Packing of pores (49)",x=0,y=0.9, just="left", gp=bold);
1078  grid.text("Hexagonal",x=0.1,y=0.82, just="left", gp=plain);
1079  grid.points(x=0.07,y=0.82,pch=16,gp=red);
1080  grid.text("Square",x=0.55,y=0.82, just="left", gp=plain);
1081  grid.points(x=0.52,y=0.82,pch=22,gp=blue);
1082  grid.text("In rows",x=0.1,y=0.4, just="left", gp=plain);
1083  grid.points(x=0.07,y=0.4,pch=23,gp=yellow);
1084  grid.text("Irregular",x=0.55,y=0.4, just="left", gp=plain);
```

Code F.2: MorphospaceFunctions.R (continued)

```
  grid.points(x=0.52,y=0.4,pch=24,gp=black);
  popViewport();
  grid.text("F", x=4/72, y=(204.5-150)/210, gp=abc, just="left");
  #And finally, X12
  par(fig=c(2/72,32/72,0,1/7),new=TRUE);
  plotMS2D(pco$points[,1:2],char="X12",m=mfull[rownames(d),],namelabs=FALSE,big=FALSE,
       axthck=axthck,axcol=axcol,bw=bw);
  vp <- viewport(x=46/72, y=0.5/7, width=30/72, height=30/210, angle=0,);
  pushViewport(vp);
  r <- readPNG("fig1labels/x12.png");
  grid.raster(r);
  grid.text("Valve face topography (10)",x=0,y=0.9, just="left", gp=bold);
  grid.text("Flat",x=0.1,y=0.82, just="left", gp=plain);
  grid.points(x=0.07,y=0.82,pch=16,gp=red);
  grid.text("Convex",x=0.55,y=0.82, just="left", gp=plain);
  grid.text("(low curvature)",x=0.55,y=0.76, just="left", gp=plain);
  grid.points(x=0.52,y=0.82,pch=22,gp=blue);
  grid.text("Convex",x=0.1,y=0.44, just="left", gp=plain);
  grid.text("(high curv.)",x=0.1,y=0.37, just="left", gp=plain);
  grid.points(x=0.07,y=0.4,pch=23,gp=yellow);
  grid.text("3 other states",x=0.55,y=0.4, just="left", gp=plain);
  grid.points(x=0.52,y=0.4,pch=24,gp=black);
  popViewport();
  grid.text("G", x=4/72, y=(204.5-180)/210, gp=abc, just="left");
  dev.off();
  #Now embed font in that file
  embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font");
}


#Function to plot the Berggren et al., 1995 Cenozoic timescale plus
#the Early and Late Cretaceous from Gradstein & Ogg, 2004 along the x-axis
#of the current plot
berg95 <- function(under=FALSE,line=1)
{
  if(under==FALSE)
  {
    bot <- par("usr")[3];
    top <- bot+(strheight("Mio",cex=0.7))*3
    txt <- top-(((top-bot)-strheight("Mio",cex=0.7))/2);
  }else{
    top <- par("usr")[3];
    bot <- top-(strheight("Mio",cex=0.7))*3
    txt <- top-(((top-bot)-strheight("Mio",cex=0.7))/2);
  }
  rect(0,bot,1.85,top, col="peachpuff",xpd=TRUE,lwd=line);
  rect(1.85,bot,5.32,top, col="khaki1",xpd=TRUE,lwd=line);
  rect(5.32,bot,23.8,top, col="yellow",xpd=TRUE,lwd=line);
  rect(23.8,bot,33.7,top, col="wheat2",xpd=TRUE,lwd=line);
  rect(33.7,bot,54.5,top, col="goldenrod1",xpd=TRUE,lwd=line);
  rect(54.5,bot,min(65,par("usr")[1]),top, col="burlywood1",xpd=TRUE,lwd=line);
  text(0.95,txt,labels="P",cex=.7,pos=1,offset=0,xpd=TRUE);
  text(3.5,txt,labels="Pl.",cex=.7,pos=1,offset=0,xpd=TRUE);
  text(14.0,txt,labels="Miocene",cex=.7,pos=1,offset=0,xpd=TRUE);
  text(28.5,txt,labels="Olig.",cex=.7,pos=1,offset=0,xpd=TRUE);
  text(44,txt,labels="Eocene",cex=.7,pos=1,offset=0,xpd=TRUE);
  text(59.7,txt,labels="Pal.",cex=.7,pos=1,offset=0,xpd=TRUE);
  if(par("usr")[1] > 65)
```

Code F.2: MorphospaceFunctions.R (continued)

```
  {
    rect(65,bot,min(99.6,par("usr")[1]),top, col="#AFD46C",xpd=TRUE,lwd=line);
    if(par("usr")[1] > 70)
    {
    text(82,txt,labels="L. Cretaceous",cex=.7,pos=1,offset=0,xpd=TRUE);
    }

  }
  if(par("usr")[1] > 99.6)
  {
    rect(99.6,bot,par("usr")[1],top, col="#94CC79",xpd=TRUE,lwd=line);
    text(108,txt,labels="E. Cret",cex=.7,pos=1,offset=0,xpd=TRUE);
  }
}

#Function to calculate convex hull volumes for a vector of time bins
#threeDonly=TRUE constrains analysis to 3D
convHullsForBins <- function(samplingmode,msls,pco,threeDonly=FALSE)
{
  #Calculate convex hull volume, for increasing # of PCO axes
  #from 3 to 10, normalizing each set of values to the highest one
  #Which dimensions to use (smallest, largest number)
  if(threeDonly == TRUE)
  {
    dimrange <- c(3,3);
  }
  else if(samplingmode == "in-bin")
  {
    dimrange <- c(3,6);
  }else if(samplingmode == "range-through")
  {
    dimrange <- c(3,10);
  }else if(samplingmode == "uw")
  {
    dimrange <- c(3,3)
  }else
  {
    dimrange <- c(3,as.numeric(min(summary(msls)[,"Length"][summary(msls)[,"Mode"] == "
         character"]))-1);
  }
  #Preallocate results variable (rows are diff. no. of PCO axes, columns are time)
  res <- matrix(NA,nrow=(dimrange[2]-dimrange[1])+1,ncol=length(msls));
  #Loop through each number of dimensions
  for(i in 1:(dim(res)[1]))
  {
    res[i,] <- convexHullVol(pco$points, msls, dim=dimrange[1]+(i-1));
    if(samplingmode != "uw")
    {
      res[i,] <- res[i,]/max(res[i,],na.rm = TRUE);
    }
  }
  return(res);
}


#Function to calculate alpha volumes for a vector of time bins
alphaVolsForBins <- function(msls,pco,alphavals)
{
```

Code F.2: MorphospaceFunctions.R (continued)

```
  #Preallocate results variable (rows are diff. alpha values, columns are time)
  resalph <- matrix(NA,nrow=length(alphavals),ncol=length(msls));
  #Loop through diff. alpha values
  for(i in 1:length(alphavals))
  {
    #Calculate alpha volumes for all time bins for given alpha value
    resalph[i,] <- alphaShapeVol(alpha=alphavals[i],points=pco$points[,1:3],msls=msls);
  }
  return(resalph);
}


#Function to plot the diversity-disparity figure
plotDivDispPubLg <- function(bins, binnames, samplingmode, N, d,sptrials=100,gentrials=10)
{
  #Load diversity functions and full (species-level) Neptune database
  source("DiversityFunctions.R");
  Nfull <- read.table(file='NeptuneProcessed.txt', header=TRUE, sep="\t");
  #Run PCO returning all PCO axes
  pco <- cmdscale(d, k=dim(d)[1]-1, eig=TRUE);
  #Ensure there are no taxa in Neptune that aren't also in morphospace
  N <- N[N$Genus %in% row.names(d),];
  #Which alpha values to use
  alphavals <- c(0.11,0.05,0.075,0.2,10);
  #Calculate disparity measures
  if(samplingmode == "in-bin" | samplingmode == "range-through")
  {
    #Get lists of genus names for each time bin
    #(be sure to have run the Neptune prep above first!!)
    msls <- getBinLists(N, bins, binnames, mode=samplingmode);
    #Calculate disparity measures for each time bin
    #Calculate mean pairwise distance
    mpwd <- meanPairwiseDist(d,msls);
    #Calculate convex hull volumes
    res <- convHullsForBins(samplingmode,msls,pco);
    #Calculate alpha shape volumes
    resalph <- alphaVolsForBins(msls,pco,alphavals);
  }else if(samplingmode == "uw") #Sampling mode is subsample by-list, unweighted
  {
    #Prep for subsampling
    intensity <- getSamplingIntensity(N=N, bins=length(binnames)-2, agemax=bins[2], agemin
        =0);
    #Choose a threshold number of lists to subsample to
    threshold <- 13;
    #Set up 3-dimensional arrays to hold results for the iterations
    hullsarr <- array(NA,dim=c(trials,10,length(binnames)));
    resalpharr <- array(NA,dim=c(trials,length(alphavals),length(binnames)));
    #And 2D-arrays for the mean distances and the genus diversities
    mpwdarr <- array(NA,dim=c(trials,length(binnames)));
    genarr <- array(NA,dim=c(trials,length(binnames)));
    #Iterate through i number of replicate subsampling trials
    for(i in 1:gentrials)
    {
      #Output which replicate is currently being run (keep track)
      cat("Morphospace trial ", i, "\n");
      #Obtain a subsampled dataset
      Nsub <- byListUWSubsample(data=N, intensity=intensity, threshold=threshold);
      #Get taxon list for this subsample
```

Code F.2: MorphospaceFunctions.R (continued)

```
      msls <- getBinLists(Nsub, bins, binnames, mode="in-bin");
      #Now get the disparity measures
      #Calculate mean pairwise distance
      mpwdarr[i,] <- meanPairwiseDist(d,msls);
      #Calculate convex hull volumes
      hulls <- convHullsForBins(samplingmode,msls,pco);
      hullsarr[i,1:dim(hulls)[1],] <- hulls;
      #Calculate alpha shape volumes
      resalpharr[i,,] <- alphaVolsForBins(msls,pco,alphavals);
      #Finally record genus diversities for each bin (used for diversity plot below)
      genuscounts <- as.numeric(summary(msls)[,"Length"]);
      names(genuscounts) <- names(msls);
      genuscounts[summary(msls)[,"Mode"] == "logical"]  <- 0;
      genarr[i,] <- genuscounts;
    }
    #Now average the values for the iterations
    mpwd <- colMeans(mpwdarr,na.rm=TRUE);
    names(mpwd) <- binnames;
    res <- colMeans(hullsarr,na.rm=TRUE);
    resalph <- colMeans(resalpharr,na.rm=TRUE);
    genuscounts <- colMeans(genarr,na.rm=TRUE);
    #Make bins with zero diversity NA (so they don't plot below)
    genuscounts[genuscounts == 0] <- NA;
    names(genuscounts) <- names(msls);
    #Calculate 95% confidence intervals on mean pairwise distances
    #Lower and upper bound for each column (=time bin)
    mpwdlowci <- array(NA,dim=dim(results)[2]);
    mpwdhici <- array(NA,dim=dim(results)[2]);
    #Loop through time bins i.e. columns
    for(a in 1:dim(mpwdarr)[2])
    {
      mpwdarr[,a] <- sort(mpwdarr[,a], na.last=TRUE);
      #If they're all NAs, then don't bother
      if(sum(!(is.na(mpwdarr[,a]))) > 0)
      {
        mpwdlowci[a] <- mpwdarr[ceiling(sum(!(is.na(mpwdarr[,a])))*0.025),a];
      }else
      {
        mpwdlowci[a] <- NA;
      }
      #If they're all NAs, then don't bother
      if(sum(!(is.na(mpwdarr[,a]))) > 0)
      {
        mpwdhici[a] <- mpwdarr[ceiling(sum(!(is.na(mpwdarr[,a])))*0.975),a];
      }else
      {
        mpwdhici[a] <- NA;
      }
    }
  }

  #Calculate diversity measures for range-through and in-bin sampling
  #Be sure to use the same taxon sampling approach as for the morphospace metrics above!
  #Obtain number of genera in morphospace time bins
  if(samplingmode == "in-bin" | samplingmode == "range-through")
  {
    genuscounts <- as.numeric(summary(msls)[,"Length"]);
    names(genuscounts) <- names(msls);
```

Code F.2: MorphospaceFunctions.R (continued)

```
    genuscounts <- genuscounts[summary(msls)[,"Mode"] != "logical"];
  }
  if(samplingmode == "range-through")
  {
    #Range-through:
    spdiv <- rangeThrough(Nfull,age.min=0,age.max=64,bins=32);
  } else if(samplingmode == "in-bin")
  {
    thing <- getSamplingIntensity(Nfull,agemin=0,agemax=64,bins=32);
    spdiv <- thing$div;
    names(spdiv) <- thing$midpoint;
  } else if(samplingmode == "uw")
  {
    intensity <- getSamplingIntensity(Nfull,agemin=0,agemax=64,bins=32);
    #Store the resulting diversity values in a matrix, the columns
    #represent the time bins, the rows represent subsamples
    results <- vector();
    #Loop through each replicate
    for (i in 1:sptrials){
      #Output which replicate is currently being run (keep track)
      cat("Sp. div. trial ", i, "\n");
      #Make a subsample (threshold has been set above in the morphospace subsampling,
      #and the same value should indeed be used for the most comparable results)
      x <- byListUWSubsample(N, intensity, threshold);
      #Calculate the diversities, etc., of the subsample
      stats <- getStats.uw(x, intensity, 8);
      #Append the diversities of the subsample to results matrix
      results <- rbind(results, stats$ns);
    }
    #Calculate average diversity curve
    spdiv <- colMeans(results);
    names(spdiv) <- seq(from=63, to=1, by=-2);
    #Calculate 95% confidence intervals
    #Lower and upper bound for each column (=time bin)
    lowci <- array(NA,dim=dim(results)[2]);
    hici <- array(NA,dim=dim(results)[2]);
    #Loop through time bins i.e. columns
    for(a in 1:dim(results)[2])
    {
      results[,a] <- sort(results[,a], na.last=TRUE);
      lowci[a] <- results[ceiling(length(!(is.na(results[,a]))))*0.025,a];
      hici[a] <- results[ceiling(length(!(is.na(results[,a]))))*0.975,a];
    }
  }


  #Plot the disparity & diversity measures
  #Get the fonts ready
  file.exists <- function( fname ) length(Sys.glob(fname))>0
  absolute.path.to.font.files <-  "/Users/bkotrc/font/";
  bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
  ## if you do not have the correct font types
  for (i in 1:length(bera.names)) {
    stopifnot( file.exists(paste(absolute.path.to.font.files, bera.names[i], ".afm", sep=""
        )) )
    stopifnot( file.exists(paste(absolute.path.to.font.files, bera.names[i], ".otf", sep=""
        )) )
  }
```

Code F.2: MorphospaceFunctions.R (continued)

```
gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files, bera.names, ".afm", sep=""))
pdfname <- "divdisp.pdf";
#Open PDF device
pdf(file=pdfname, bg="white", width=(14.8/cm(1)), height=(21.5/cm(1)), pointsize=10, family=gillsans, colormodel="cmyk");
axthck <- 0.3;
par(mfrow=c(5,1), mar=c(1.5,4.1,1.5,4.1), oma=c(15.1,7.5,0,7.5), lwd=axthck, font.main=1, cex.main=1.5);

#Plot mean pairwise distance
if(samplingmode == "uw")
{
  plot(names(mpwd),mpwd,axes=FALSE,bty="n",type="n",lwd=axthck,xlim=c(112,0),ylim=c(min(mpwdlowci,na.rm=TRUE)-diff(range(c(mpwdlowci,mpwdhici),na.rm=TRUE))*.18,max(mpwdhici,na.rm=TRUE)),ylab="Mean pairwise distance",xlab="");
}else
{
  plot(names(mpwd),mpwd,axes=FALSE,bty="n",type="n",xlim=c(112,0),ylim=c(min(mpwd,na.rm=TRUE)-diff(range(mpwd,na.rm=TRUE))*.18,max(mpwd,na.rm=TRUE)),ylab="Mean pairwise distance",xlab="");
}
axis(1,lwd=axthck);
axis(2,lwd=axthck);
title(main="A",adj=0.05,line=-1);
points(names(mpwd)[1:2],mpwd[1:2],pch=16,cex=0.8);
points(names(mpwd)[3:length(mpwd)],mpwd[3:length(mpwd)],pch=16,type="o",cex=0.8,lty=1,lwd=1);
berg95(line=axthck);
#Add error bars if mode is "uw"
if(samplingmode == "uw")
{
  arrows(as.numeric(names(mpwd)),mpwd,as.numeric(names(mpwd)),mpwdlowci,length=0,lwd=0.25);
  arrows(as.numeric(names(mpwd)),mpwd,as.numeric(names(mpwd)),mpwdhici,length=0,lwd=0.25);
}

#Plot convex hull volumes
#UW subsampled volumes come back unstandardized, so standardize them
if(samplingmode == "uw")
{
  for(i in 1:dim(res)[1])
  {
    res[i,] <- res[i,]/max(res[i,],na.rm = TRUE);
  }
}
#Set up axes, plot no points
plot(names(mpwd),res[1,],bty="n",type="n",xlim=c(112,0),ylim=c(-0.15,1),ylab="Normalized convex hull volume",xlab="",axes=FALSE);
axis(2, lwd=axthck);
axis(1, lwd=axthck);
title(main="B",adj=0.05,line=-1);
#Plot results for 4-10D, grey lines, Cenozoic
for(i in 2:(dim(res)[1])){
  points(names(mpwd)[3:length(mpwd)],res[i,3:length(mpwd)],type="l",col="grey",lwd=1);
}
#Results for 4-10D, grey circles, Mesozoic
```

Code F.2: MorphospaceFunctions.R (continued)

```
  points(rep(names(mpwd)[1],times=dim(res)[1]-1),res[2:dim(res)[1],1],col="grey",cex=0.8);
  points(rep(names(mpwd)[2],times=dim(res)[1]-1),res[2:dim(res)[1],2],col="grey",cex=0.8);
  #Results for 3D, filled circles
  points(names(mpwd)[1:2],res[1,1:2],pch=16,cex=0.8);
  points(names(mpwd)[3:length(mpwd)],res[1,3:length(mpwd)],pch=16,type="o",cex=0.8,lwd=1);
  #Labels for curves
  if(samplingmode != "uw")
  {
    text(35,0.2,labels=paste("PC 1-",dim(res)[1]+2,sep=""),cex=0.8);
  }
  if(samplingmode == "uw")
  {
    text(52,0.9, labels="PC 1-3",cex=0.8);
  }else
  {
    text(40,0.9, labels="PC 1-3",cex=0.8);
  }
  berg95(line=axthck);

  #Plot alpha shape results
  #Set up axes, plot no points
  plot(names(mpwd),resalph[1,],bty="n",axes=FALSE,type="n",xlim=c(112,0),ylim=c((min(
       resalph,na.rm=TRUE)-(diff(range(resalph,na.rm=TRUE))*0.15)),max(resalph,na.rm=TRUE))
       ,ylab="Alpha shape volume",xlab="");
  axis(2, lwd=axthck);
  axis(1, lwd=axthck);
  title(main="C",adj=0.05,line=-1);
  #Plot results for other alpha values, grey lines, Cenozoic
  for(i in 2:dim(resalph)[1]){
    points(names(mpwd)[3:length(mpwd)],resalph[i,3:length(mpwd)],type="l",col="grey",lwd=1)
        ;
  }
  #Results for other alpha values, grey circles, Mesozoic
  points(rep(names(mpwd)[1],times=dim(resalph)[1]-1),resalph[2:dim(resalph)[1],1],col="grey
       ",cex=0.8);
  points(rep(names(mpwd)[2],times=dim(resalph)[1]-1),resalph[2:dim(resalph)[1],2],col="grey
       ",cex=0.8);
  #Results for alpha=0.11, filled circles
  points(names(mpwd)[1:2],resalph[1,1:2],pch=16,cex=0.8);
  points(names(mpwd)[3:length(mpwd)],resalph[1,3:length(mpwd)],pch=16,type="o",cex=0.8,lwd
       =1);
  #Add labels to show alpha values for the curves (top one nudged up)
  text(names(mpwd[length(mpwd)]),resalph[length(alphavals),length(mpwd)]+0.00075,labels=
       bquote(alpha == .(alphavals[length(alphavals)])),cex=0.8,pos=4,xpd=TRUE);
  for(j in 1:(length(alphavals)-1))
  {
    text(names(mpwd[length(mpwd)]),resalph[j,length(mpwd)],labels=bquote(alpha == .(
        alphavals[j])),cex=0.8,pos=4,xpd=TRUE);
  }
  #text(names(mpwd[length(mpwd)]),resalph[length(alphavals),length(mpwd)]+0.0005,labels=
       paste("A=",alphavals[length(alphavals)],sep=""),cex=0.8,pos=4,xpd=TRUE);
  #text(names(mpwd[length(mpwd)]),resalph[1:(length(alphavals)-1),length(mpwd)],labels=
       paste("A=",alphavals[1:(length(alphavals)-1)],sep=""),cex=0.8,pos=4,xpd=TRUE);
  #Timescale
  berg95(line=axthck);

  #Plot alpha volume per genus
  ct <- genuscounts;
```

Code F.2: MorphospaceFunctions.R (continued)

```
  for (i in 1:(length(alphavals)-1))
  {
    ct <- rbind(ct,genuscounts)
  }
  alphpergen <- resalph/ct;
  #Set up axes, plot no points
  plot(names(mpwd),alphpergen[1,],axes=FALSE,bty="n",type="n",xlim=c(112,0),ylim=c((min(
       alphpergen,na.rm=TRUE)-(diff(range(alphpergen,na.rm=TRUE))*0.15)),max(alphpergen,na.
       rm=TRUE)),ylab="Alpha shape volume per genus",xlab="");
  axis(1, lwd=axthck);
  axis(2, lwd=axthck);
  title(main="D",adj=0.05,line=-1);
  #Plot results for other alpha values, grey lines, Cenozoic
  for(i in 2:dim(resalph)[1]){
    points(names(mpwd)[3:length(mpwd)],alphpergen[i,3:length(mpwd)],type="l",col="grey",lwd
         =1);
  }
  #Results for other alpha values, grey circles, Mesozoic
  points(rep(names(mpwd)[1],times=dim(resalph)[1]-1),alphpergen[2:dim(resalph)[1],1],col="
       grey",cex=0.8);
  points(rep(names(mpwd)[2],times=dim(resalph)[1]-1),alphpergen[2:dim(resalph)[1],2],col="
       grey",cex=0.8);
  #Results for alpha=0.11, filled circles
  points(names(mpwd)[1:2],alphpergen[1,1:2],pch=16,cex=0.8);
  points(names(mpwd)[3:length(mpwd)],alphpergen[1,3:length(mpwd)],pch=16,type="o",cex=0.8,
       lwd=1);
  #Add labels to show alpha values for the curves (top one nudged up)
  text(names(mpwd[length(mpwd)]),alphpergen[length(alphavals),length(mpwd)]+0.00002,labels=
       bquote(alpha == .(alphavals[length(alphavals)])),cex=0.8,pos=4,xpd=TRUE);
  text(names(mpwd[length(mpwd)]),alphpergen[length(alphavals),length(mpwd)]+0.000001,labels
       =bquote(alpha == .(alphavals[length(alphavals)-1])),cex=0.8,pos=4,xpd=TRUE);
  for(j in 1:(length(alphavals)-2))
  {
    text(names(mpwd[length(mpwd)]),alphpergen[j,length(mpwd)],labels=bquote(alpha == .(
         alphavals[j])),cex=0.8,pos=4,xpd=TRUE);
  }
  #text(names(mpwd[length(mpwd)]),resalph[length(alphavals),length(mpwd)]+0.0005,labels=
       paste("A=",alphavals[length(alphavals)],sep=""),cex=0.8,pos=4,xpd=TRUE);
  #text(names(mpwd[length(mpwd)]),resalph[1:(length(alphavals)-1),length(mpwd)],labels=
       paste("A=",alphavals[1:(length(alphavals)-1)],sep=""),cex=0.8,pos=4,xpd=TRUE);
  #Timescale
  berg95(line=axthck);

  #Finally, add plot showing diversity--at both species and genus levels
  #Plot # of genera in morphospace
  plot(names(genuscounts)[3:length(genuscounts)],genuscounts[3:length(genuscounts)],bty="n"
       ,yaxt="n",type="o",lwd=1,axes=FALSE,xlim=c(112,0),ylim=c((min(genuscounts,na.rm=TRUE
       )-(diff(range(genuscounts,na.rm=TRUE))*0.2)),max(genuscounts,na.rm=TRUE)+(diff(range
       (genuscounts,na.rm=TRUE))*0.1)),ylab="",pch=15,col="grey",xpd=TRUE,cex=0.8);
  axis(side=4,col="grey", lwd=axthck);
  title(main="E",adj=0.05,line=-4);
  mtext("Genus richness",side=4,line=3,cex=par("cex"));
  points(names(genuscounts)[1:2],genuscounts[1:2],cex=0.8,pch=15,col="grey");
  #plot(names(genuscounts),genuscounts,xaxt="n",yaxt="n",bty="n",xlim=c(112,0),ylim=c((min(
       genuscounts,na.rm=TRUE)-(diff(range(genuscounts,na.rm=TRUE))*0.25)),max(genuscounts,
       na.rm=TRUE)),ylab="",xlab="",pch=16,col="grey",xpd=TRUE);
  mtext(text="Geologic Time (Ma)",side=1,line=3,xpd=TRUE,cex=par("cex"));
  #Overplot # of species in Neptune on same subplot
```

Code F.2: MorphospaceFunctions.R (continued)

```
  par(new=TRUE);
  plot(names(spdiv),spdiv,xaxt="n",yaxt="n",type="o",bty="n",lwd=1,xlim=c(112,0),ylim=c((
       min(spdiv,na.rm=TRUE)-(diff(range(spdiv,na.rm=TRUE))*0.25)),max(spdiv,na.rm=TRUE)),
       ylab="",xlab="",pch=16,xpd=TRUE,cex=0.8);
  #plot(names(spdiv),spdiv,bty="n",xlim=c(112,0),ylim=c((min(spdiv,na.rm=TRUE)-(diff(range(
       spdiv,na.rm=TRUE))*0.25)),max(spdiv,na.rm=TRUE)),ylab="Species richness (Neptune)",
       pch=16,xpd=TRUE);
  axis(2, lwd=axthck);
  axis(1, lwd=axthck);
  mtext("Species richness ",side=2,line=3,cex=par("cex"));
  #Timescale
  berg95(line=axthck);
  #Add legend
  legend(x="topleft",inset=0.05,bty="n",lwd=1,pt.cex=0.8, cex=0.8, legend=c("Species
       richness (Neptune)", "Genus richness (morphospace)"),col=c("black","grey"),pch=c
       (16,15));
  #Add error bars if mode is "uw"
  if(samplingmode == "uw")
  {
    arrows(as.numeric(names(spdiv)),spdiv,as.numeric(names(spdiv)),lowci,length=0,lwd=0.25)
        ;
    arrows(as.numeric(names(spdiv)),spdiv,as.numeric(names(spdiv)),hici,length=0,lwd=0.25);
  }

  #Close graphics device
  dev.off();

  #Now embed font in that file
  embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font");

  #Return mean pairwise distances, 3D convex hull volumes, alpha volumes, and genus counts
  results <- cbind(mpwd,res[1,],resalph[1,],genuscounts);
  colnames(results)[2:3] <- c("chullvols","alphavols");
  return(as.data.frame(results));
}


#Function to plot the diversity-disparity figure for paper 2 (subsampling)
plotDivDispPub2Lg <- function(bins, binnames, samplingmode, N, d,sptrials=100,gentrials=10,
     sendback=FALSE)
{
  #Load diversity functions and full (species-level) Neptune database
  source("DiversityFunctions.R");
  Nfull <- read.table(file='NeptuneProcessed.txt', header=TRUE, sep="\t");
  #Run PCO returning all PCO axes
  pco <- cmdscale(d, k=dim(d)[1]-1, eig=TRUE);
  #Ensure there are no taxa in Neptune that aren't also in morphospace
  N <- N[N$Genus %in% row.names(d),];
  #Which alpha values to use
  alphavals <- c(0.11,0.05,0.075,0.2,10);

  #Calculate disparity measures
  if(samplingmode == "in-bin" | samplingmode == "range-through")
  {
    #Get lists of genus names for each time bin
    #(be sure to have run the Neptune prep above first!!)
    msls <- getBinLists(N, bins, binnames, mode=samplingmode);
    #Calculate disparity measures for each time bin
```

Code F.2: MorphospaceFunctions.R (continued)

```
    #Calculate mean pairwise distance
    mpwd <- meanPairwiseDist(d,msls);
    #Calculate convex hull volumes
    res <- convHullsForBins(samplingmode,msls,pco);
    #Calculate alpha shape volumes
    resalph <- alphaVolsForBins(msls,pco,alphavals);
  }else if(samplingmode == "uw" | samplingmode == "cr" | samplingmode == "ow" |
       samplingmode == "sqs")
  #Sampling mode is subsampling
  {
    #Prep for subsampling
    intensity <- getSamplingIntensity(N=N, bins=length(binnames), agemax=64, agemin=0);
    #Choose a threshold number of lists to subsample to for UW
    threshold <- 13;
    #Choose a threshold number of occurrences to subsample to for OW, CR
    thresholdocc <- 100;
    #Choose a shareholder quorum for SQS
    q <- 0.5;
    #Set up 3-dimensional array to hold results for the alpha volume iterations
    resalpharr <- array(NA,dim=c(gentrials,length(alphavals),length(binnames)));
    #And 2D-arrays for the mean distances, hull volumes, and genus diversities
    mpwdarr <- array(NA,dim=c(gentrials,length(binnames)));
    hullsarr <- array(NA,dim=c(gentrials,length(binnames)));
    genarr <- array(NA,dim=c(gentrials,length(binnames)));
    #Iterate through i number of replicate subsampling trials
    for(i in 1:gentrials)
    {
      #Output which replicate is currently being run (keep track)
      cat("Morphospace trial ", i, "\n");
      #Obtain a subsampled dataset
      if(samplingmode == "uw")
      {
        Nsub <- byListUWSubsample(data=N, intensity=intensity, threshold=threshold);
      }
      if(samplingmode == "ow")
      {
        Nsub <- owSubsample(data=N, intensity=intensity, threshold=thresholdocc);
      }
      if(samplingmode == "sqs")
      {
        Nsub <- sqsSubsample(data=N, intensity=intensity, q=q);
      }
      else{
        Nsub <- naiveRarefactionSubsample(data=N, intensity=intensity, threshold=
             thresholdocc);
        }
      #Get taxon list for this subsample
      msls <- getBinLists(Nsub, bins, binnames, mode="in-bin");
      #Now get the disparity measures
      #Calculate mean pairwise distance
      mpwdarr[i,] <- meanPairwiseDist(d,msls);
      #Calculate convex hull volumes (in 3D only)
      hulls <- convexHullVol(p=pco$points,msls=msls,dim=3);
      hullsarr[i,] <- hulls;
      #Calculate alpha shape volumes
      resalpharr[i,,] <- alphaVolsForBins(msls,pco,alphavals);
      #Finally record genus diversities for each bin (used for diversity plot below)
      genuscounts <- as.numeric(summary(msls)[,"Length"]);
```

Code F.2: MorphospaceFunctions.R (continued)

```
      names(genuscounts) <- names(msls);
      genuscounts[summary(msls)[,"Mode"] == "logical"]  <- 0;
      genarr[i,] <- genuscounts;
    }
    #Now average the values for the iterations
    mpwd <- colMeans(mpwdarr,na.rm=TRUE);
    names(mpwd) <- binnames;
    res <- colMeans(hullsarr,na.rm=TRUE);
    resalph <- colMeans(resalpharr,na.rm=TRUE);
    genuscounts <- colMeans(genarr,na.rm=TRUE);
    hullsav <- colMeans(hullsarr,na.rm=TRUE);
    names(hullsav) <- binnames;
    #Make bins with zero diversity NA (so they don't plot below)
    genuscounts[genuscounts == 0] <- NA;
    names(genuscounts) <- names(msls);
    #Calculate 95% confidence intervals on mean pairwise distances
    #Lower and upper bound for each column (=time bin)
    mpwdlowci <- array(NA,dim=length(mpwd));
    mpwdhici <- array(NA,dim=length(mpwd));
    #Loop through time bins i.e. columns
    for(a in 1:dim(mpwdarr)[2])
    {
      mpwdarr[,a] <- sort(mpwdarr[,a], na.last=TRUE);
      #If they're all NAs, then don't bother
      if(sum(!(is.na(mpwdarr[,a]))) > 0)
      {
        mpwdlowci[a] <- mpwdarr[ceiling(sum(!(is.na(mpwdarr[,a])))*0.025),a];
      }else
      {
        mpwdlowci[a] <- NA;
      }
      #If they're all NAs, then don't bother
      if(sum(!(is.na(mpwdarr[,a]))) > 0)
      {
        mpwdhici[a] <- mpwdarr[ceiling(sum(!(is.na(mpwdarr[,a])))*0.975),a];
      }else
      {
        mpwdhici[a] <- NA;
      }
    }
    #Calculate 95% confidence intervals on convex hull volumes
    #Lower and upper bound for each column (=time bin)
    hulllowci <- array(NA,dim=length(mpwd));
    hullhici <- array(NA,dim=length(mpwd));
    #Loop through time bins i.e. columns
    for(a in 1:dim(hullsarr)[2])
    {
      hullsarr[,a] <- sort(hullsarr[,a], na.last=TRUE);
      #If they're all NAs, then don't bother
      if(sum(!(is.na(hullsarr[,a]))) > 0)
      {
        hulllowci[a] <- hullsarr[ceiling(sum(!(is.na(hullsarr[,a])))*0.025),a];
      }else
      {
        hulllowci[a] <- NA;
      }
      #If they're all NAs, then don't bother
      if(sum(!(is.na(hullsarr[,a]))) > 0)
```

Code F.2: MorphospaceFunctions.R (continued)

```
    {
      hullhici[a] <- hullsarr[ceiling(sum(!(is.na(hullsarr[,a])))*0.975),a];
    }else
    {
      hullhici[a] <- NA;
    }
  }
  #Calculate 95% confidence intervals on 0.11 alpha volumes
  #Lower and upper bound for each column (=time bin)
  alphalowci <- array(NA,dim=length(mpwd));
  alphahici <- array(NA,dim=length(mpwd));
  #Make life easier by extracting just the 0.11 alpha part of resalpharr
  resalpharr11 <- resalpharr[,1,];
  #Loop through time bins i.e. columns
  for(a in 1:dim(resalpharr11)[2])
  {
    resalpharr11[,a] <- sort(resalpharr11[,a], na.last=TRUE);
    #If they're all NAs, then don't bother
    if(sum(!(is.na(resalpharr11[,a]))) > 0)
    {
      alphalowci[a] <- resalpharr11[ceiling(sum(!(is.na(resalpharr11[,a])))*0.025),a];
    }else
    {
      alphalowci[a] <- NA;
    }
    #If they're all NAs, then don't bother
    if(sum(!(is.na(resalpharr11[,a]))) > 0)
    {
      alphahici[a] <- resalpharr11[ceiling(sum(!(is.na(resalpharr11[,a])))*0.975),a];
    }else
    {
      alphahici[a] <- NA;
    }
  }
}
#Calculate diversity measures for range-through and in-bin sampling
#Be sure to use the same taxon sampling approach as for the morphospace metrics above!
#Obtain number of genera in morphospace time bins
if(samplingmode == "in-bin" | samplingmode == "range-through")
{
  genuscounts <- as.numeric(summary(msls)[,"Length"]);
  names(genuscounts) <- names(msls);
  #This line was a mistake:
  #genuscounts <- genuscounts[summary(msls)[,"Mode"] != "logical"];
  #This should work better:
  #Include bins with zero occurrences
  genuscounts[summary(msls)[,"Mode"] == "logical"]  <- 0;
  #But mark them as NA, so they won't plot
  genuscounts[genuscounts == 0] <- NA;
}
if(samplingmode == "range-through")
{
  #Range-through:
  spdiv <- rangeThrough(Nfull,age.min=0,age.max=64,bins=32);
  #Make zero-occurrence bins NA so they don't plot
  spdiv[spdiv==0] <- NA;
} else if(samplingmode == "in-bin")
{
```

Code F.2: MorphospaceFunctions.R (continued)

```
    thing <- getSamplingIntensity(Nfull,agemin=0,agemax=64,bins=32);
    spdiv <- thing$div;
    names(spdiv) <- thing$midpoint;
    #Make zero-occurrence bins NA so they don't plot
    spdiv[spdiv==0] <- NA;
  } else if(samplingmode == "uw")
  {
    intensity <- getSamplingIntensity(Nfull,agemin=0,agemax=64,bins=32);
    #Store the resulting diversity values in a matrix, the columns
    #represent the time bins, the rows represent subsamples
    results <- vector();
    #Loop through each replicate
    for (i in 1:sptrials){
      #Output which replicate is currently being run (keep track)
      cat("Sp. div. trial ", i, "\n");
      #Make a subsample (threshold has been set above in the morphospace subsampling,
      #and the same value should indeed be used for the most comparable results)
      x <- byListUWSubsample(Nfull, intensity, threshold);
      #Calculate the diversities, etc., of the subsample
      stats <- getStats.uw(x, intensity, 8);
      #Append the diversities of the subsample to results matrix
      results <- rbind(results, stats$ns);
    }
    #Calculate average diversity curve
    spdiv <- colMeans(results);
    names(spdiv) <- seq(from=63, to=1, by=-2);
    #Calculate 95% confidence intervals
    #Lower and upper bound for each column (=time bin)
    lowci <- array(NA,dim=dim(results)[2]);
    hici <- array(NA,dim=dim(results)[2]);
    #Loop through time bins i.e. columns
    for(a in 1:dim(results)[2])
    {
      results[,a] <- sort(results[,a], na.last=TRUE);
      lowci[a] <- results[ceiling(length(!(is.na(results[,a])))*0.025),a];
      hici[a] <- results[ceiling(length(!(is.na(results[,a])))*0.975),a];
    }
    #Hack bug fix: remove points with zero diversity so they don't plot
    hici[spdiv == 0] <- NA;
    lowci[spdiv == 0] <- NA;
    spdiv[spdiv == 0] <- NA;
  } else if(samplingmode == "cr")
  {
    intensity <- getSamplingIntensity(Nfull,agemin=0,agemax=64,bins=32);
    #Store the resulting diversity values in a matrix, the columns
    #represent the time bins, the rows represent subsamples
    results <- vector();
    #Loop through each replicate
    for (i in 1:sptrials){
      #Output which replicate is currently being run (keep track)
      cat("Sp. div. trial ", i, "\n");
      #Make a subsample (threshold has been set above in the morphospace subsampling,
      #and the same value should indeed be used for the most comparable results)
      x <- naiveRarefactionSubsample(Nfull, intensity, thresholdocc);
      #Calculate the diversities, etc., of the subsample
      stats <- getStats.uw(x, intensity, 8);
      #Append the diversities of the subsample to results matrix
      results <- rbind(results, stats$ns);
```

Code F.2: MorphospaceFunctions.R (continued)

```
    }
    #Calculate average diversity curve
    spdiv <- colMeans(results);
    names(spdiv) <- seq(from=63, to=1, by=-2);
    #Calculate 95% confidence intervals
    #Lower and upper bound for each column (=time bin)
    lowci <- array(NA,dim=dim(results)[2]);
    hici <- array(NA,dim=dim(results)[2]);
    #Loop through time bins i.e. columns
    for(a in 1:dim(results)[2])
    {
      results[,a] <- sort(results[,a], na.last=TRUE);
      lowci[a] <- results[ceiling(length(!(is.na(results[,a])))*0.025),a];
      hici[a] <- results[ceiling(length(!(is.na(results[,a])))*0.975),a];
    }
    #Hack bug fix: remove points with zero diversity so they don't plot
    hici[spdiv == 0] <- NA;
    lowci[spdiv == 0] <- NA;
    spdiv[spdiv == 0] <- NA;
  } else if(samplingmode == "ow")
  {
    intensity <- getSamplingIntensity(Nfull,agemin=0,agemax=64,bins=32);
    #Store the resulting diversity values in a matrix, the columns
    #represent the time bins, the rows represent subsamples
    results <- vector();
    #Loop through each replicate
    for (i in 1:sptrials){
      #Output which replicate is currently being run (keep track)
      cat("Sp. div. trial ", i, "\n");
      #Make a subsample (threshold has been set above in the morphospace subsampling,
      #and the same value should indeed be used for the most comparable results)
      x <- owSubsample(Nfull, intensity, thresholdocc);
      #Calculate the diversities, etc., of the subsample
      stats <- getStats.uw(x, intensity, 8);
      #Append the diversities of the subsample to results matrix
      results <- rbind(results, stats$ns);
    }
    #Calculate average diversity curve
    spdiv <- colMeans(results);
    names(spdiv) <- seq(from=63, to=1, by=-2);
    #Calculate 95% confidence intervals
    #Lower and upper bound for each column (=time bin)
    lowci <- array(NA,dim=dim(results)[2]);
    hici <- array(NA,dim=dim(results)[2]);
    #Loop through time bins i.e. columns
    for(a in 1:dim(results)[2])
    {
      results[,a] <- sort(results[,a], na.last=TRUE);
      lowci[a] <- results[ceiling(length(!(is.na(results[,a])))*0.025),a];
      hici[a] <- results[ceiling(length(!(is.na(results[,a])))*0.975),a];
    }
    #Hack bug fix: remove points with zero diversity so they don't plot
    hici[spdiv == 0] <- NA;
    lowci[spdiv == 0] <- NA;
    spdiv[spdiv == 0] <- NA;
  } else if(samplingmode == "sqs")
  {
    intensity <- getSamplingIntensity(Nfull,agemin=0,agemax=64,bins=32);
```

Code F.2: MorphospaceFunctions.R (continued)

```
1845    #Store the resulting diversity values in a matrix, the columns
1846    #represent the time bins, the rows represent subsamples
1847    results <- vector();
1848    #Loop through each replicate
1849    for (i in 1:sptrials){
1850      #Output which replicate is currently being run (keep track)
1851      cat("Sp. div. trial ", i, "\n");
1852      #Make a subsample (threshold has been set above in the morphospace subsampling,
1853      #and the same value should indeed be used for the most comparable results)
1854      x <- sqsSubsample(Nfull, intensity, q);
1855      #Calculate the diversities, etc., of the subsample
1856      stats <- getStats.uw(x, intensity, 8);
1857      #Append the diversities of the subsample to results matrix
1858      results <- rbind(results, stats$ns);
1859    }
1860    #Calculate average diversity curve
1861    spdiv <- colMeans(results);
1862    names(spdiv) <- seq(from=63, to=1, by=-2);
1863    #Calculate 95% confidence intervals
1864    #Lower and upper bound for each column (=time bin)
1865    lowci <- array(NA,dim=dim(results)[2]);
1866    hici <- array(NA,dim=dim(results)[2]);
1867    #Loop through time bins i.e. columns
1868    for(a in 1:dim(results)[2])
1869    {
1870      results[,a] <- sort(results[,a], na.last=TRUE);
1871      lowci[a] <- results[ceiling(length(!(is.na(results[,a])))*0.025),a];
1872      hici[a] <- results[ceiling(length(!(is.na(results[,a])))*0.975),a];
1873    }
1874    #Hack bug fix: remove points with zero diversity so they don't plot
1875    hici[spdiv == 0] <- NA;
1876    lowci[spdiv == 0] <- NA;
1877    spdiv[spdiv == 0] <- NA;
1878  }
1879  #Plot the disparity & diversity measures
1880  #Get the fonts ready
1881  file.exists <- function( fname ) length(Sys.glob(fname))>0
1882  absolute.path.to.font.files <-  "/Users/bkotrc/font/";
1883  bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
1884  ## if you do not have the correct font types
1885  for (i in 1:length(bera.names)) {
1886    stopifnot( file.exists(paste(absolute.path.to.font.files, bera.names[i], ".afm", sep=""
          )) )
1887    stopifnot( file.exists(paste(absolute.path.to.font.files, bera.names[i], ".otf", sep=""
          )) )
1888  }
1889  gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files, bera.names, ".afm",
       sep=""))
1890  pdfname <- paste("divdisp","-",samplingmode,"-",gentrials,"-trials",".pdf",sep="");
1891  #Open PDF device
1892  pdf(file=pdfname, bg="white", width=(14.8/cm(1)), height=(21.5/cm(1)), pointsize=10,
       family=gillsans, colormodel="cmyk");
1893  axthck <- 0.3;
1894  par(mfrow=c(5,1), mar=c(1.5,4.1,1.5,4.1), oma=c(15.1,7.5,0,7.5), lwd=axthck, font.main=1,
        cex.main=1.5);
1895
1896  #Plot mean pairwise distance
```

Code F.2: MorphospaceFunctions.R (continued)

```
1897   if(samplingmode == ”uw” | samplingmode == ”cr” | samplingmode == ”ow” | samplingmode == ”
           sqs”)
1898   {
1899 plot(names(mpwd),mpwd,axes=FALSE,bty=”n”,type=”n”,lwd=axthck,xlim=c(65,0),ylim=c(min(
       mpwdlowci,na.rm=TRUE)-diff(range(c(mpwdlowci,mpwdhici),na.rm=TRUE))*.18,max(mpwdhici,
       na.rm=TRUE)),ylab=”Mean pairwise distance”,xlab=””);
1900   }else
1901   {
1902 plot(names(mpwd),mpwd,axes=FALSE,bty=”n”,type=”n”,xlim=c(65,0),ylim=c(min(mpwd,na.rm=TRUE)-
       diff(range(mpwd,na.rm=TRUE))*.23,(max(mpwd,na.rm=TRUE)*1.05)),ylab=”Mean pairwise
       distance”,xlab=””);
1903   }
1904   axis(1,lwd=axthck);
1905   axis(2,lwd=axthck);
1906   title(main=”A”,adj=0.05,line=-1);
1907   points(names(mpwd),mpwd,pch=16,type=”o”,cex=0.8,lty=1, lwd=1);
1908   berg95(line=axthck);
1909   #Add error bars if mode is ”uw”
1910   if(samplingmode == ”uw” | samplingmode == ”cr” | samplingmode == ”ow” | samplingmode == ”
           sqs”)
1911   {
1912 arrows(as.numeric(names(mpwd)),mpwd,as.numeric(names(mpwd)),mpwdlowci,length=0,lwd=0.25);
1913 arrows(as.numeric(names(mpwd)),mpwd,as.numeric(names(mpwd)),mpwdhici,length=0,lwd=0.25);
1914   }
1915   #Plot convex hull volumes
1916   #Two whole different blocks of plot code follow--in-bin/RT vs. subsampling
1917   #First: if sampling is in-bin or RT
1918   if(samplingmode == ”in-bin” | samplingmode == ”range-through”)
1919   {
1920   #Set up axes, plot no points
1921 plot(names(mpwd),res[1,],bty=”n”,type=”n”,xlim=c(65,0),ylim=c(-0.15,1),ylab=”Normalized
       convex hull volume”,xlab=””,axes=FALSE);
1922   axis(2, lwd=axthck);
1923   axis(1, lwd=axthck);
1924   title(main=”B”,adj=0.05,line=-1);
1925   #Plot results for 4-10D, grey lines, Cenozoic
1926   for(i in 2:(dim(res)[1])){
1927     points(names(mpwd),res[i,],type=”l”,col=”grey”,lwd=1);
1928   }
1929   #Results for 3D, filled circles
1930   points(names(mpwd),res[1,],pch=16,type=”o”,cex=0.8,lwd=1);
1931   #Labels for curves
1932   if(samplingmode != ”uw”)
1933   {
1934     text(35,0.2,labels=paste(”PC 1-”,dim(res)[1]+2,sep=””),cex=0.8);
1935   }
1936   if(samplingmode == ”uw” | samplingmode == ”cr” | samplingmode == ”ow” | samplingmode == ”
           sqs”)
1937   {
1938     text(52,0.9, labels=”PC 1-3”,cex=0.8);
1939   }else
1940   {
1941     text(40,0.9, labels=”PC 1-3”,cex=0.8);
1942   }
1943   berg95(line=axthck);
1944   }else
1945   #Second: if sampling mode is subsampling
1946   {
```

Code F.2: MorphospaceFunctions.R (continued)

```
    #Set up axes, plot no points
plot(names(mpwd),hullsav,bty="n",type="n",xlim=c(65,0),ylim=c(min(hulllowci,na.rm=TRUE)-
     diff(range(c(hulllowci,hullhici),na.rm=TRUE))*.18,max(hullhici,na.rm=TRUE)),ylab="
     Convex hull volume",xlab="",axes=FALSE);
  axis(2, lwd=axthck);
  axis(1, lwd=axthck);
  title(main="B",adj=0.05,line=-1);
  #Results for 3D, filled circles
  points(names(mpwd),hullsav,pch=16,type="o",cex=0.8,lwd=1);
  #Labels for curves
  if(samplingmode != "uw")
  {
    text(35,0.2,labels=paste("PC 1-",dim(res)[1]+2,sep=""),cex=0.8);
  }
  if(samplingmode == "uw" | samplingmode == "cr" | samplingmode == "ow"| samplingmode == "
       sqs")
  {
    text(52,0.9, labels="PC 1-3",cex=0.8);
  }else
  {
    text(40,0.9, labels="PC 1-3",cex=0.8);
  }
  berg95(line=axthck);
  #Add error bars in subsampling mode
  if(samplingmode == "uw" | samplingmode == "cr" | samplingmode == "ow" | samplingmode == "
       sqs")
  {
arrows(as.numeric(names(mpwd)),hullsav,as.numeric(names(mpwd)),hulllowci,length=0,lwd=0.25)
     ;
arrows(as.numeric(names(mpwd)),hullsav,as.numeric(names(mpwd)),hullhici,length=0,lwd=0.25);
  }
  }
  #Plot alpha shape results
  #Set up axes, plot no points
plot(names(mpwd),resalph[1,],bty="n",axes=FALSE,type="n",xlim=c(65,0),ylim=c((min(resalph,
     na.rm=TRUE)-(diff(range(resalph,na.rm=TRUE))*0.15)),max(resalph,na.rm=TRUE)),ylab="
     Alpha shape volume",xlab="");
  axis(2, lwd=axthck);
  axis(1, lwd=axthck);
  title(main="C",adj=0.05,line=-1);
  #Plot results for other alpha values, grey lines, Cenozoic
  for(i in 2:dim(resalph)[1]){
    points(names(mpwd),resalph[i,],type="l",col="grey",lwd=1);
  }
  #Results for alpha=0.11, filled circles
  points(names(mpwd),resalph[1,],pch=16,type="o",cex=0.8,lwd=1);
  #Add labels to show alpha values for the curves (top one nudged up)
  text(names(mpwd[length(mpwd)]),resalph[length(alphavals),length(mpwd)]+0.00075,labels=
       bquote(alpha == .(alphavals[length(alphavals)])),cex=0.8,pos=4,xpd=TRUE);
  for(j in 1:(length(alphavals)-1))
  {
    text(names(mpwd[length(mpwd)]),resalph[j,length(mpwd)],labels=bquote(alpha == .(
         alphavals[j])),cex=0.8,pos=4,xpd=TRUE);
  }
  #Timescale
  berg95(line=axthck);
  #Add error bars in subsampling mode
```

Code F.2: MorphospaceFunctions.R (continued)

```
  if(samplingmode == "uw" | samplingmode == "cr" | samplingmode == "ow" | samplingmode == "sqs")
  {
arrows(as.numeric(names(mpwd)),resalph[1,],as.numeric(names(mpwd)),alphalowci,length=0,lwd=0.25);
arrows(as.numeric(names(mpwd)),resalph[1,],as.numeric(names(mpwd)),alphahici,length=0,lwd=0.25);
  }
  #Plot alpha volume per genus
  ct <- genuscounts;
  for (i in 1:(length(alphavals)-1))
  {
    ct <- rbind(ct,genuscounts)
  }
  alphpergen <- resalph/ct;
  #Set up axes, plot no points
plot(names(mpwd),alphpergen[1,],axes=FALSE,bty="n",type="n",xlim=c(65,0),ylim=c((min(alphpergen,na.rm=TRUE)-(diff(range(alphpergen,na.rm=TRUE))*0.15)),max(alphpergen,na.rm=TRUE)),ylab="Alpha shape volume per genus",xlab="");
  axis(1, lwd=axthck);
  axis(2, lwd=axthck);
  title(main="D",adj=0.05,line=-1);
  #Plot results for other alpha values, grey lines, Cenozoic
  for(i in 2:dim(resalph)[1]){
    points(names(mpwd),alphpergen[i,],type="l",col="grey",lwd=1);
  }
  points(names(mpwd),alphpergen[1,],pch=16,type="o",cex=0.8,lwd=1);
  #Add labels to show alpha values for the curves (top one nudged up)
  text(names(mpwd[length(mpwd)]),alphpergen[length(alphavals),length(mpwd)]+0.00002,labels=bquote(alpha == .(alphavals[length(alphavals)])),cex=0.8,pos=4,xpd=TRUE);
  text(names(mpwd[length(mpwd)]),alphpergen[length(alphavals),length(mpwd)]+0.000001,labels=bquote(alpha == .(alphavals[length(alphavals)-1])),cex=0.8,pos=4,xpd=TRUE);
  for(j in 1:(length(alphavals)-2))
  {
    text(names(mpwd[length(mpwd)]),alphpergen[j,length(mpwd)],labels=bquote(alpha == .(alphavals[j])),cex=0.8,pos=4,xpd=TRUE);
  }
  #Timescale
  berg95(line=axthck);
  #Finally, add plot showing diversity--at both species and genus levels
  #Plot # of genera in morphospace
plot(names(genuscounts),genuscounts,bty="n",yaxt="n",type="o",lwd=1,axes=FALSE,xlim=c(65,0),ylim=c((min(genuscounts,na.rm=TRUE)-(diff(range(genuscounts,na.rm=TRUE))*0.2)),max(genuscounts,na.rm=TRUE)+(diff(range(genuscounts,na.rm=TRUE))*0.1)),ylab="",pch=15,col="grey",xpd=TRUE,cex=0.8);
  axis(side=4,col="grey", lwd=axthck);
  title(main="E",adj=0.05,line=-4);
  mtext("Genus richness",side=4,line=3,cex=par("cex"));
  mtext(text="Geologic Time (Ma)",side=1,line=3,xpd=TRUE,cex=par("cex"));
  #Overplot # of species in Neptune on same subplot
  par(new=TRUE);
#If subsampling, take error bars into account
if(samplingmode == "cr" | samplingmode == "uw" | samplingmode == "ow" | samplingmode == "sqs")
{
plot(names(spdiv),spdiv,xaxt="n",yaxt="n",type="o",bty="n",lwd=1,xlim=c(65,0),ylim=c((min(lowci,na.rm=TRUE)-(diff(range(c(lowci,hici),na.rm=TRUE))*0.25)),max(hici,na.rm=TRUE)),ylab="",xlab="",pch=16,xpd=TRUE,cex=0.8);
```

Code F.2: MorphospaceFunctions.R (continued)

```
  }else
  #If not subsampling, forget about error bars in scaling the plot area
  {   plot(names(spdiv),spdiv,xaxt="n",yaxt="n",type="o",bty="n",lwd=1,xlim=c(65,0),ylim=c
       ((min(spdiv,na.rm=TRUE)-(diff(range(spdiv,na.rm=TRUE))*0.25)),max(spdiv,na.rm=TRUE))
       ,ylab="",xlab="",pch=16,xpd=TRUE,cex=0.8);
  }
  axis(2, lwd=axthck);
  axis(1, lwd=axthck);
  mtext("Species richness ",side=2,line=3,cex=par("cex"));
  #Timescale
  berg95(line=axthck);
  #Add legend
  legend(x="topleft",inset=0.05,bty="n",lwd=1,pt.cex=0.8, cex=0.8, legend=c("Species
       richness (Neptune)", "Genus richness (morphospace)"),col=c("black","grey"),pch=c
       (16,15));
  #Add error bars if mode is "uw"
  if(samplingmode == "uw" | samplingmode == "cr" | samplingmode == "ow" | samplingmode == "
       sqs")
  {
arrows(as.numeric(names(spdiv)),spdiv,as.numeric(names(spdiv)),lowci,length=0,lwd=0.25);
arrows(as.numeric(names(spdiv)),spdiv,as.numeric(names(spdiv)),hici,length=0,lwd=0.25);
  }
  #Close graphics device
  dev.off();
  #Now embed font in that file
  embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font");
  #Return values---currently only works for in-bin/range-through
  if(sendback==TRUE)
  {
    results <- cbind(mpwd,res[1,],resalph[1,]);
    colnames(results)[2:3] <- c("chullvols","alphavols");
    rownames(results) <- names(mpwd);
    return(as.data.frame(results));
  }
}


#Function to calculate diversity-disparity metrics for paper 2 "data culling"
#Copy of plotDivDispPub2Lg with the plotting bits excised,
#and the alpha shape volume calculations removed
#Also, removed the subsampling modes (not needed for "data culling" comparisons),
#and all the diversity calculations. So, way stripped down
plotDivDispPub2LgDQ <- function(bins, binnames, samplingmode, N, d,sptrials=100,gentrials
     =10,sendback=FALSE)
{
  #Load diversity functions and full (species-level) Neptune database
  source("DiversityFunctions.R");
  #Run PCO returning all PCO axes
  pco <- cmdscale(d, k=dim(d)[1]-1, eig=TRUE);
  #Ensure there are no taxa in Neptune that aren't also in morphospace
  N <- N[N$Genus %in% row.names(d),];
  #Calculate disparity measures
  if(samplingmode == "in-bin" | samplingmode == "range-through")
  {
    #Get lists of genus names for each time bin
    msls <- getBinLists(N, bins, binnames, mode=samplingmode);
    #Calculate disparity measures for each time bin
    #Calculate mean pairwise distance
```

Code F.2: MorphospaceFunctions.R (continued)

```
    mpwd <- meanPairwiseDist(d,msls);
    #Calculate convex hull volumes
    #(Lying about samplingmode in order to force use of only 3-6
    #PCO dimensions, to avoid problems with low-diversity bins)
    res <- convHullsForBins("in-bin",msls,pco);
  }
  #Return values---currently only works for in-bin/range-through
  if(sendback==TRUE)
  {
    results <- cbind(mpwd,res[1,]);
    colnames(results)[2] <- "chullvols";
    rownames(results) <- names(mpwd);
    return(as.data.frame(results));
  }
}


#Function to calculate the average convex hull volume per list
#Returns averages, and top and bottom of 50% confidence interval:
#$mean - mean alpha shape volume
#$reshici - top of 50% of volumes (25th percentile)
#$resloci - bottom of 50% of volumes (75th percentile)
getAlphaDispVol <- function(bins, binnames,N, d)
{
  #Load diversity functions and full (species-level) Neptune database
  source("DiversityFunctions.R");
  #Run PCO returning all PCO axes
  pco <- cmdscale(d, k=dim(d)[1]-1, eig=TRUE);
  points <- pco$points;
  #Ensure there are no taxa in Neptune that aren't also in morphospace
  N <- N[N$Genus %in% row.names(d),];
  #Preallocate variable to hold average for each bin
  mean <- vector(mode="numeric",length=length(binnames));
  #Preallocate variable to hold top of confidence interval
  reshici <- vector(mode="numeric",length=length(binnames));
  #Preallocate variable to hold bottom of confidence interval
  resloci <- vector(mode="numeric",length=length(binnames));
  #Preallocate variable to hold list length average for each bin
  divresults <- vector(mode="numeric",length=length(binnames));
  #Go through time bins
  #First (oldest) time bin
  if(length(unique(N[N$Sample.Age > bins[1],]$Genus)) == 0)
  {
    results[1] <- NA;
  }else
  {
    #Subset N to this time bin
    Nsub <- N[N$Sample.Age > bins[1],];
    #Lists in this time bin
    lists <- unique(Nsub$List);
    #Preallocate variable to hold alpha shape volumes for each list
    binresults <- vector(mode="numeric",length=length(lists));
    binresultsdiv <- vector(mode="numeric",length=length(lists));
    #Go through each list
    for(j in 1:length(lists))
    {
      #Subset Nsub to jth list
      NList <- Nsub[Nsub$List == lists[j],];
```

Code F.2: MorphospaceFunctions.R (continued)

```
    #Genus names in jth list
    genera <- unique(NList$Genus);
    #Alpha volume occupied by jth list
    #Extract relevant part of p, the PCO points matrix
    pi <- points[genera,1:3];
    #Diversity (list length)
    binresultsdiv[j] <- length(genera);
    #Conv. hull volume for those points
    #The convex hull volume calculation appears to fail with fewer than 4
    #genera on a list, so calculate only for lists with more than 3
    if(length(genera) > 3)
    {
      #Conv. hull volume for those points
      binresults[j] <- convhulln(pi, options="FA")[[3]];
      #Get the alpha shape for the selected point cloud as an "ashape3d" class object
      #alphaobj <- ashape3d(pi,alpha=0.11);
      #Calculate volume of the alpha shape
      #binresults[j] <- volume_ashape3d(alphaobj);
    }else{
      binresults[j] <- NA;
    }

  }
  #Now calculate average volume
  mean[1] <- mean(binresults,na.rm=TRUE);
  #Confidence intervals
  #First, sort
  binresults <- sort(binresults, na.last=TRUE);
  resloci[1] <- binresults[ceiling(sum(!(is.na(binresults)))*0.25)];
  reshici[1] <- binresults[ceiling(sum(!(is.na(binresults)))*0.75)];

  #Average list length
  divresults[1] <- mean(binresultsdiv);
}
#Loop through middle time bins
for(i in 1:(length(bins)-1))
{
  #What if the time bin is empty?
  if(length(unique(N[(N$Sample.Age <= bins[i] & N$Sample.Age > bins[i+1]),]$Genus)) == 0)
  {
    results[i+1] <- NA;
  }else
  {
    #Subset N to this time bin
    Nsub <- N[N$Sample.Age <= bins[i] & N$Sample.Age > bins[i+1],];
    #Lists in this time bin
    lists <- unique(Nsub$List);
    #Preallocate variable to hold alpha shape volumes for each list
    binresults <- vector(mode="numeric",length=length(lists));
    binresultsdiv <- vector(mode="numeric",length=length(lists));
    #Go through each list
    for(j in 1:length(lists))
    {
      #Subset Nsub to jth list
      NList <- Nsub[Nsub$List == lists[j],];
      #Genus names in jth list
      genera <- unique(NList$Genus);
      #Alpha volume occupied by jth list
```

Code F.2: MorphospaceFunctions.R (continued)

```
      #Extract relevant part of p, the PCO points matrix
      pi <- points[genera,1:3];
      #Diversity (list length)
      binresultsdiv[j] <- length(genera);
      #The convex hull volume calculation appears to fail with fewer than 4
      #genera on a list, so calculate only for lists with more than 3
      if(length(genera) > 3)
      {
        #Conv. hull volume for those points
        binresults[j] <- convhulln(pi, options="FA")[[3]];
        #Get the alpha shape for the selected point cloud as an "ashape3d" class object
        #alphaobj <- ashape3d(pi,alpha=0.11);
        #Calculate volume of the alpha shape
        #binresults[j] <- volume_ashape3d(alphaobj);
      }else{
        binresults[j] <- NA;
      }
    }
    #Now calculate average volume
    mean[i+1] <- mean(binresults,na.rm=TRUE);
    #Confidence intervals
    binresults <- sort(binresults, na.last=TRUE);
    resloci[i+1] <- binresults[ceiling(sum(!(is.na(binresults)))*0.25)];
    reshici[i+1] <- binresults[ceiling(sum(!(is.na(binresults)))*0.75)];

    divresults[i+1] <- mean(binresultsdiv);
  }
}
#Last (youngest) time bin
#Subset N to this time bin
Nsub <- N[N$Sample.Age <= bins[length(bins)],];
#Lists in this time bin
lists <- unique(Nsub$List);
#Preallocate variable to hold alpha shape volumes for each list
binresults <- vector(mode="numeric",length=length(lists));
binresultsdiv <- vector(mode="numeric",length=length(lists));
#Go through each list
for(j in 1:length(lists))
{
  #Subset Nsub to jth list
  NList <- Nsub[Nsub$List == lists[j],];
  #Genus names in jth list
  genera <- unique(NList$Genus);
  #Alpha volume occupied by jth list
  #Extract relevant part of p, the PCO points matrix
  pi <- points[genera,1:3];
  #Diversity (list length)
  binresultsdiv[j] <- length(genera);
  #The convex hull volume calculation appears to fail with fewer than 4
  #genera on a list, so calculate only for lists with more than 3
  if(length(genera) > 3)
  {
    #Conv. hull volume for those points
    binresults[j] <- convhulln(pi, options="FA")[[3]];
    #Get the alpha shape for the selected point cloud as an "ashape3d" class object
    #alphaobj <- ashape3d(pi,alpha=0.11);
    #Calculate volume of the alpha shape
    #binresults[j] <- volume_ashape3d(alphaobj);
```

Code F.2: MorphospaceFunctions.R (continued)

```
    }else{
      binresults[j] <- NA;
    }
  }
  #Now calculate average volume
  mean[length(mean)] <- mean(binresults,na.rm=TRUE);
  #Confidence intervals
  binresults <- sort(binresults, na.last=TRUE);
  resloci[length(mean)] <- binresults[ceiling(sum(!(is.na(binresults)))*0.25)];
  reshici[length(mean)] <- binresults[ceiling(sum(!(is.na(binresults)))*0.75)];
  divresults[length(divresults)] <- mean(binresultsdiv);
  #Send anything back?
  results <- as.data.frame(cbind(mean,reshici,resloci),row.names=as.character(binnames));
  return(results);
}


#Function to calculate the average mean pairwise distance per list
#Returns averages, and top and bottom of 95% confidence interval:
#$mean - mean alpha shape volume
#$reshici - top of 50% of means (25th percentile)
#$resloci - bottom of 50% of means (75th percentile)
getAlphaDispMPWD <- function(bins, binnames,N, d)
{
  #Load diversity functions and full (species-level) Neptune database
  source("DiversityFunctions.R");
  #Run PCO returning all PCO axes
  pco <- cmdscale(d, k=dim(d)[1]-1, eig=TRUE);
  points <- pco$points;
  #Ensure there are no taxa in Neptune that aren't also in morphospace
  N <- N[N$Genus %in% row.names(d),];
  #Preallocate variable to hold average for each bin
  mean <- vector(mode="numeric",length=length(binnames));
  #Preallocate variable to hold top of confidence interval
  reshici <- vector(mode="numeric",length=length(binnames));
  #Preallocate variable to hold bottom of confidence interval
  resloci <- vector(mode="numeric",length=length(binnames));
  #Preallocate variable to hold list length average for each bin
  divresults <- vector(mode="numeric",length=length(binnames));
  #Go through time bins
  #First (oldest) time bin
  if(length(unique(N[N$Sample.Age > bins[1],]$Genus)) == 0)
  {
    results[1] <- NA;
  }else
  {
    #Subset N to this time bin
    Nsub <- N[N$Sample.Age > bins[1],];
    #Lists in this time bin
    lists <- unique(Nsub$List);
    #Preallocate variable to hold alpha shape volumes for each list
    binresults <- vector(mode="numeric",length=length(lists));
    binresultsdiv <- vector(mode="numeric",length=length(lists));
    #Go through each list
    for(j in 1:length(lists))
    {
      #Subset Nsub to jth list
      NList <- Nsub[Nsub$List == lists[j],];
```

Code F.2: MorphospaceFunctions.R (continued)

```
      #Genus names in jth list
      genera <- unique(NList$Genus);
      #Diversity (list length)
      binresultsdiv[j] <- length(genera);
      #Need 2 taxa in a list to get a distance
      if(length(genera) > 2)
      {
        #MPWD for those points
        #Extract relevant part of d matrix
          di <- d[genera,genera];
          #Get distance matrix as lower triangular
          dlower <- di[lower.tri(di)];
          #Now get mean of all values in lower triangular
          binresults[j] <- mean(dlower);
      }else{
        binresults[j] <- NA;
      }
    }
    #Now calculate average
    mean[1] <- mean(binresults,na.rm=TRUE);
    #Confidence intervals
    #First, sort
    binresults <- sort(binresults, na.last=TRUE);
    resloci[1] <- binresults[ceiling(sum(!(is.na(binresults)))*0.25)];
    reshici[1] <- binresults[ceiling(sum(!(is.na(binresults)))*0.75)];
    #Average list length
    divresults[1] <- mean(binresultsdiv);
  }
  #Loop through middle time bins
  for(i in 1:(length(bins)-1))
  {
    #What if the time bin is empty?
    if(length(unique(N[(N$Sample.Age <= bins[i] & N$Sample.Age > bins[i+1]),]$Genus)) == 0)
    {
      results[i+1] <- NA;
    }else
    {
      #Subset N to this time bin
      Nsub <- N[N$Sample.Age <= bins[i] & N$Sample.Age > bins[i+1],];
      #Lists in this time bin
      lists <- unique(Nsub$List);
      #Preallocate variable to hold alpha shape volumes for each list
      binresults <- vector(mode="numeric",length=length(lists));
      binresultsdiv <- vector(mode="numeric",length=length(lists));
      #Go through each list
    for(j in 1:length(lists))
    {
      #Subset Nsub to jth list
      NList <- Nsub[Nsub$List == lists[j],];
      #Genus names in jth list
      genera <- unique(NList$Genus);
      #Diversity (list length)
      binresultsdiv[j] <- length(genera);
      #Need 2 taxa in a list to get a distance
      if(length(genera) > 2)
      {
        #MPWD for those points
        #Extract relevant part of d matrix
```

Code F.2: MorphospaceFunctions.R (continued)

```
          di <- d[genera,genera];
          #Get distance matrix as lower triangular
          dlower <- di[lower.tri(di)];
          #Now get mean of all values in lower triangular
          binresults[j] <- mean(dlower);
      }else{
        binresults[j] <- NA;
      }
    }
      #Now calculate average
      mean[i+1] <- mean(binresults,na.rm=TRUE);
      #Confidence intervals
      binresults <- sort(binresults, na.last=TRUE);
      resloci[i+1] <- binresults[ceiling(sum(!(is.na(binresults)))*0.25)];
      reshici[i+1] <- binresults[ceiling(sum(!(is.na(binresults)))*0.75)];

      divresults[i+1] <- mean(binresultsdiv);
    }
  }
  #Last (youngest) time bin
  #Subset N to this time bin
  Nsub <- N[N$Sample.Age <= bins[length(bins)],];
  #Lists in this time bin
  lists <- unique(Nsub$List);
  #Preallocate variable to hold alpha shape volumes for each list
  binresults <- vector(mode="numeric",length=length(lists));
  binresultsdiv <- vector(mode="numeric",length=length(lists));
  #Go through each list
    for(j in 1:length(lists))
    {
      #Subset Nsub to jth list
      NList <- Nsub[Nsub$List == lists[j],];
      #Genus names in jth list
      genera <- unique(NList$Genus);
      #Diversity (list length)
      binresultsdiv[j] <- length(genera);
      #Need 2 taxa in a list to get a distance
      if(length(genera) > 2)
      {
        #MPWD for those points
        #Extract relevant part of d matrix
          di <- d[genera,genera];
          #Get distance matrix as lower triangular
          dlower <- di[lower.tri(di)];
          #Now get mean of all values in lower triangular
          binresults[j] <- mean(dlower);
      }else{
        binresults[j] <- NA;
      }
    } #Now calculate average
  mean[length(mean)] <- mean(binresults,na.rm=TRUE);
  #Confidence intervals
  binresults <- sort(binresults, na.last=TRUE);
  resloci[length(mean)] <- binresults[ceiling(sum(!(is.na(binresults)))*0.25)];
  reshici[length(mean)] <- binresults[ceiling(sum(!(is.na(binresults)))*0.75)];
  divresults[length(divresults)] <- mean(binresultsdiv);
  #Send anything back?
  results <- as.data.frame(cbind(mean,reshici,resloci),row.names=as.character(binnames));
```

Code F.2: MorphospaceFunctions.R (continued)

```
2439   return(results);
2440 }
2441
2442
2443 #Function to plot results of "alpha disparity" analysis
2444 plotAlphaDisparity <- function(alphadispv,alphadispd)
2445 {
2446   #Get the fonts ready
2447   file.exists <- function( fname ) length(Sys.glob(fname))>0
2448   absolute.path.to.font.files <-  "/Users/bkotrc/font/";
2449   bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
2450   ## if you do not have the correct font types
2451   for (i in 1:length(bera.names)) {
2452     stopifnot( file.exists(paste(absolute.path.to.font.files, bera.names[i], ".afm", sep=""
           )) )
2453     stopifnot( file.exists(paste(absolute.path.to.font.files, bera.names[i], ".otf", sep=""
           )) )
2454   }
2455   gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files, bera.names, ".afm",
         sep=""))
2456   pdfname <- "alphadisp.pdf";
2457   #Open PDF device
2458   pdf(file=pdfname, bg="white", width=(14.8/cm(1)), height=(8/cm(1)), pointsize=7, family=
         gillsans, colormodel="cmyk");
2459   #Now plot it up
2460   par(mfrow=c(2,1), mar=c(1.5,4.1,1.5,4.1), oma=c(3.1,7.5,0,7.5));
2461 axthck <- 0.3;
2462   #Plot volume disparity (chulls)
2463   alphadisp <- alphadispv
2464   alphadisp[alphadisp==0] <- NA;
2465   plot(as.numeric(rownames(alphadisp)),alphadisp$mean,xlim=c(max(as.numeric(rownames(
         alphadisp))),0),ylim=(c(min(alphadisp$resloci,na.rm=TRUE)-0.001,max(alphadisp$
         reshici,na.rm=TRUE))),xlab="Age",ylab="Conv. hull vol. per list",type="n",bty="n");
2466   #Confidence intervals
2467   arrows(as.numeric(rownames(alphadisp)),alphadisp$resloci,as.numeric(rownames(alphadisp)),
         alphadisp$reshici,length=0,lwd=0.25);
2468   points(as.numeric(rownames(alphadisp)),alphadisp$mean,pch=16,type="o")
2469   berg95();
2470   title(main="A",adj=0.05,line=-1, cex.main=1.5, font.main=1);
2471   #Plot distance disparity (mpwd)
2472   alphadisp <- alphadispd
2473   alphadisp[alphadisp==0] <- NA;
2474   plot(as.numeric(rownames(alphadisp)),alphadisp$mean,xlim=c(max(as.numeric(rownames(
         alphadisp))),0),ylim=(c(min(alphadisp$resloci,na.rm=TRUE),max(alphadisp$reshici,na.
         rm=TRUE))),xlab="Age",ylab="Mean pairwise dist. per list",type="n",bty="n");
2475   #Confidence intervals
2476   arrows(as.numeric(rownames(alphadisp)),alphadisp$resloci,as.numeric(rownames(alphadisp)),
         alphadisp$reshici,length=0,lwd=0.25);
2477   points(as.numeric(rownames(alphadisp)),alphadisp$mean,pch=16,type="o")
2478   berg95();
2479   mtext(text="Geologic Time (Ma)",side=1,line=3,xpd=TRUE,cex=par("cex"));
2480   title(main="B",adj=0.05,line=-1, cex.main=1.5, font.main=1);
2481   dev.off();
2482 }
2483
2484
2485
```

Code F.2: MorphospaceFunctions.R (continued)

```
#Function to calculate the % of genera in each time bin with a specified character state of interest
#relative to other (specified) character states
#Function takes:
#chint - string with the character state of interest, referring to a column in the mstates matrix
#choth - vector of strings with the other character states of that character, referring to columns in the mstates matrix
#mstates - boolean matrix of m genera by n character states (the m matrix resolved into states)
#msls - list of vectors of strings, containing the genus names present for each time bin
#Function returns:
#pctchint - vector of %ages, representing the % of genera in each time bin with the character of interest
getCharPct <- function(chint, choth, mstates, msls)
{
  #Set up vector to carry the counts for the character states for each bin
  nchint <- vector(mode="numeric", length=length(msls));
  nchoth <- vector(mode="numeric", length=length(msls));
  #Go through each time bin
  for(i in 1:length(msls))
  {
    #Subset mstates for current bin taxa
    mstcur <- mstates[msls[[i]],];
    #How many taxa have the state of interest?
    nchint[i] <- sum(mstcur[,chint]);
    #How many taxa have the other states?
    nchoth[i] <- sum(mstcur[,choth]);
  }
  return(100*(nchint/(nchint+nchoth)));
}

#Function to plot the % of genera in each time bin with a specified character state of interest
#Takes:
#pctchint - returned by getCharPct() function
#binnames - numeric vector containing mid-point ages of time bins corresponding to pctchint
#chintname, chothname (optional) - text labels to be plotted in over the top and bottom of the graph
#                                 explaining the character(s) chosen
#... - ellipsis parameter used to pass other parameters to plot
#Returns nothing (calls base plot function so will produce graphic output as product)
plotCharPct <- function(pctchint,binnames,chintname="",chothname="",...)
{
  par(lwd=0.3,bg=NA);
  plot(binnames,pctchint,xlim=c(633,0),ylim=c(0,100),type="l",xaxs="i",yaxs="i",ylab=expression(paste(bar("%"), " of genera with states")),xlab="",xpd=TRUE,xaxt="n",yaxt="n",lwd=0.3,...);
  axis(side=2, line=0,lwd=0.3);
mtext(expression(paste(bar("%"), " of genera with states")), side=2, line=4.5,cex=par("cex"));
  polygon(c(binnames[1],binnames,0,0),c(0,pctchint,pctchint[length(pctchint)],0),col="darkgrey");
  berg95(under=TRUE,line=0.3);
  text(0.5,10,chintname,col="white",pos=2,cex=1,font=1);
  text(0.5,90,chothname,col="black",pos=2,cex=1,font=1);
}

```

Code F.2: MorphospaceFunctions.R (continued)

```
#Function to plot the ordination method sensitivity analysis
#figure for paper 2 (subsampling), comparing NMDS and PCO results
#(Only works for samplingmode="in-bin" or "range-through")
plotOrdinationComparison <- function(bins, binnames, samplingmode="in-bin", N, d)
{
  #Load diversity functions and full (species-level) Neptune database
  source("DiversityFunctions.R");
  Nfull <- read.table(file='NeptuneProcessed.txt', header=TRUE, sep="\t");
  #Run PCO returning all PCO axes
  pco <- cmdscale(d, k=dim(d)[1]-1, eig=TRUE);
  #Run NMDS algorithm returning 2 axes
  nmds <- isoMDS(d, y=cmdscale(d,3), k=3);
  #Ensure there are no taxa in Neptune that aren't also in morphospace
  N <- N[N$Genus %in% row.names(d),];
  #Which alpha values to use
  alphavals <- c(0.11,0.05,0.075,0.2,10);

  #Calculate disparity measures
  if(samplingmode == "in-bin" | samplingmode == "range-through")
  {
    #Get lists of genus names for each time bin
    #(be sure to have run the Neptune prep above first!!)
    msls <- getBinLists(N, bins, binnames, mode=samplingmode);
    #Calculate disparity measures for each time bin
    #Calculate mean pairwise distance
    mpwd <- meanPairwiseDist(d,msls);
    #Calculate convex hull volumes
    res <- convHullsForBins(samplingmode,msls,pco);
    res2 <- convHullsForBins(samplingmode,msls,nmds,threeDonly=TRUE);
    #Calculate alpha shape volumes
    resalph <- alphaVolsForBins(msls,pco,alphavals);
    resalph2 <- alphaVolsForBins(msls,nmds,alphavals);
  }

  #Plot the disparity measures
  #Get the fonts ready
  file.exists <- function( fname ) length(Sys.glob(fname))>0
  absolute.path.to.font.files <-  "/Users/bkotrc/font/";
  bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
  ## if you do not have the correct font types
  for (i in 1:length(bera.names)) {
    stopifnot( file.exists(paste(absolute.path.to.font.files, bera.names[i], ".afm", sep=""
        )) )
    stopifnot( file.exists(paste(absolute.path.to.font.files, bera.names[i], ".otf", sep=""
        )) )
  }
  gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files, bera.names, ".afm",
      sep=""))
  pdfname <- "ordinationcomparison.pdf"
  #Open PDF device
  pdf(file=pdfname, bg="white", width=(22.2/cm(1)), height=(14.8/cm(1)), pointsize=10,
      family=gillsans, colormodel="cmyk");
  axthck <- 0.3;
  layout(matrix(c(1,2,3,4), 2, 2, byrow = TRUE),widths=c(1.5,1))
  par(mar=c(4.1,4.1,1.5,4.1),oma=c(9.9,7.5,0,7.5), lwd=axthck, font.main=1, cex.main=1.5);
  #Define colors for PCO and NMDS in plots
  pcol <- "royalblue"
  ncol <- "firebrick"
```

Code F.2: MorphospaceFunctions.R (continued)

```
  #Plot chull results
  if(samplingmode == "in-bin" | samplingmode == "range-through")
  {
  #Set up axes, plot no points
plot(names(mpwd),res[1,],bty="n",type="n",xlim=c(65,0),ylim=c(-0.15,1),ylab="Normalized
    convex hull volume",xlab="Geologic Time (Ma)",axes=FALSE);
  axis(2, lwd=axthck);
  axis(1, lwd=axthck);
  title(main="A",adj=0.05,line=-1);
  #Results for 3D, filled circles
  points(names(mpwd),res[1,],pch=16,type="o",cex=0.8,lwd=1,col=pcol);
  points(names(mpwd),res2[1,],pch=15,type="o",cex=0.8,lwd=1,col=ncol);
  berg95(line=axthck);
legend(x=15,y=0.3,legend=c("PCO","NMDS"),col=c(pcol,ncol),pch=c(16,15),box.lwd=axthck,lty="
    solid",pt.cex=.8,cex=0.75)
  }
  #Plot comparison of the two (colors for xlab, ylab)!
  plot(res[1,],res2[1,],type="n",xlab="",ylab="",bty="n",axes=FALSE,xaxs = "i",yaxs="i",
      xlim=c(0,1),ylim=c(0,1))
  title(main="B",adj=0.05,line=-1);
  mtext("PCO-based volume",side=1,cex=0.8,line=3,col=pcol)
  mtext("NMDS-based volume",side=2,cex=0.8,line=3,col=ncol)
  axis(2, lwd=axthck);
  axis(1, lwd=axthck);
  #points(x=c(0,1),y=c(0,1),type="l",lty=2,lwd=axthck)
  points(res[1,],res2[1,],pch=3,lwd=0.8)
  #Linear model (regression)
  results <- data.frame(res[1,],res2[1,])
  names(results) <- c("pco","nmds")
  model <- lm(nmds ~ pco,data = results)
  abline(model,lty=2,lwd=axthck)
  text(0.8,0.2,labels=bquote(R^2 == .(round(summary(model)$r.squared,digits=2))),font=3)
  #Plot alpha shape results
  #Set up axes, plot no points
plot(names(mpwd),resalph[1,],bty="n",axes=FALSE,type="n",xlim=c(65,0),ylim=c((min(resalph,
    na.rm=TRUE)-(diff(range(resalph,na.rm=TRUE))*0.15)),max(resalph,na.rm=TRUE)),ylab="
    Alpha shape volume",xlab="Geologic Time (Ma)");
  axis(2, lwd=axthck);
  axis(1, lwd=axthck);
  title(main="C",adj=0.05,line=-1);
  #Results for alpha=0.11, filled circles (PCO)
  points(names(mpwd),resalph[1,],pch=16,type="o",cex=0.8,lwd=1,col=pcol);
  #Results for alpha=0.11, filled circles (NMDS)
  points(names(mpwd),resalph2[1,],pch=15,type="o",cex=0.8,lwd=1,col=ncol);
  #Timescale
  berg95(line=axthck);
  #X-axis label
legend(x=15,y=0.003,legend=c("PCO","NMDS"),col=c(pcol,ncol),pch=c(16,15),box.lwd=axthck,lty
    ="solid",pt.cex=.8,cex=0.75)
  #Plot comparison of the two (colors for xlab, ylab)!
  plot(resalph[1,],resalph2[1,],type="n",xlab="",ylab="",bty="n",axes=FALSE,xaxs = "i",yaxs
      ="i",xlim=c(0,.01),ylim=c(0,.01))
  axis(2, lwd=axthck);
  axis(1, lwd=axthck);
  title(main="D",adj=0.05,line=-1);
  mtext("PCO-based volume",side=1,cex=0.8,line=3,col=pcol)
  mtext("NMDS-based volume",side=2,cex=0.8,line=3,col=ncol)
  #points(x=c(0,1),y=c(0,1),type="l",lty=2,lwd=axthck)
```

Code F.2: MorphospaceFunctions.R (continued)

```
  points(resalph[1,],resalph2[1,],pch=3,lwd=0.8)
  #Linear model (regression)
  results2 <- data.frame(resalph[1,],resalph2[1,])
  names(results2) <- c("pco","nmds")
  model2 <- lm(nmds ~ pco,data = results2)
  abline(model2,lty=2,lwd=axthck)
  text(0.008,0.002,font=3,labels=bquote(R^2 == .(round(summary(model2)$r.squared,digits=2))
       ))
  #Close graphics device
  dev.off();
  #Now embed font in that file
  embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font");
}


#Function to plot the data quality comparison figure
#for paper 2 (subsampling), comparing results from 80% and 100% thresholds
#(Only works for samplingmode="in-bin" or "range-through")
#Produces a figure that looks just like the one from plotOrdinationComparison()
plotDQ80100Comparison <- function()
{
  #Options: "in-bin", "range-through","uw","cr","sqs"
samplingmode <- "in-bin";
#(N.b.: because 65 myrs is an odd number, the youngest interval is only 1 myr long)
bins <- c(seq(from=62,to=2,by=-2));
#Set time bin names
binnames <- c(seq(from=64,to=2,by=-2)-1);
#Read in the raw data matrix
mfull <- read.table(file='Matrix.txt', header=TRUE, sep=",", row.names=1,colClasses="
     character")
#Read in Neptune database, with Cretaceous occurrences added, and Genus name field,
#and with typos/mistakes/synonyms corrected
N <- read.table(file='NeptuneGenNamesCorrCret.txt', header=TRUE, sep="\t",as.is=TRUE)
#For paper 2 (subsampling), throw out the added Cretaceous stuff---Neptune only
N <- N[N$Sample.Age < 65,]
#Assess the completeness (genera and characters) of the matrix
dq <- getDataQuality(mfull, inv="?");
#Characters range from 67%-100% complete, genera from 57%
#So, start at 50% or better, go to 100%
#complevs <- c(50,60,70,80,90,100);
#Or go through each level that's in the data
complevs <- sort(unique(c(floor(unique(dq$gen)),floor(unique(dq$char)))))
#For each completeness level, we're going to want an r^2 value
#for each disparity metric
#Ready vector to hold them
mpwd <- vector("numeric",length=length(complevs));
cvol <- mpwd;
rsq <- as.data.frame(cbind(mpwd,cvol));
rownames(rsq) <- complevs;
pvals <- rsq;
#Which taxon sampling mode to use?
#Options: "in-bin", "range-through","uw","cr","sqs"
samplingmode <- "in-bin";
#(N.b.: because 65 myrs is an odd number, the youngest interval is only 1 myr long)
bins <- c(seq(from=62,to=2,by=-2));
#Set time bin names
binnames <- c(seq(from=64,to=2,by=-2)-1);
#Read distance matrix based on data culled to >80% using "?" only as missing
```

Code F.2: MorphospaceFunctions.R (continued)

```
#(as used in paper 1)
dref <- as.matrix(read.table(file='DistanceMatrixCulled8080?nosingles.txt', header=TRUE, row.names=1, sep=",",as.is=TRUE))
#We're going to need a reference set of disparity results
#to compare to---using the 80% culling threshold from paper 1
dispref <- plotDivDispPub2LgDQ(bins, binnames, samplingmode, N, dref, sptrials=1,gentrials=1,sendback=TRUE)
i <- 32
#Make a copy of the full morphospace matrix for culling
  mcull <- mfull;
  #Keep track of the size of the matrix, to stop when it's stable
  msize <- c(1000,1000);
  while(!identical(dim(mcull),msize)){
    #Update matrix size
    msize <- dim(mcull)
    #Cull by ith completeness level percentage
    #mcull <- cullMatrix(mcull,getDataQuality(mcull, inv="?"),complevs[i],complevs[i]);
    #Characters=100
    mcull <- cullMatrix(mcull,getDataQuality(mcull, inv="?"),0,complevs[i]);
    #Genus=100
    mcull <- cullMatrix(mcull,getDataQuality(mcull, inv="?"),complevs[i],0);
    #Now throw out "uninformative" characters with less than two
    #states having one or more valid entries
    uninf <- vector(length=ncol(mcull));
    names(uninf) <- colnames(mcull);
    for(j in 1:ncol(mcull))
    {
      #If the character has less than 2 states with more than 1 valid entry
      if(sum(table(as.factor(makeNumeric(mcull[,j]))) > 1) < 2)
      {
        #Flag as uninformative
        uninf[j] <- TRUE;
      }
    }
    #Cull the matrix to take out those "uninformative" characters
    mcull <- mcull[,!uninf];
  }
  #Now that we have a culled matrix, let's run the disparity metrics
  #First, get distance matrix
  dcull <- getDistMatrix(mcull);
  #Throw out Neptune occurrences that aren't in the matrix
  #(this step is crucial or it will break)
  Ncull <- N[N$Genus %in% row.names(dcull),];
  #Now run in-bin disparity metrics using this matrix
  dispcull <- plotDivDispPub2LgDQ(bins, binnames, samplingmode, Ncull, dcull, sptrials=1000,gentrials=1000,sendback=TRUE);
  mpwd <- dispref$mpwd
  mpwd2 <- dispcull$mpwd
  res <- dispref$chullvols
  res <- rbind(res,res)
  res2 <- dispcull$chullvols
  res2 <- rbind(res2,res2)
  names(mpwd) <- binnames
  #Plot the disparity measures
  #Get the fonts ready
  file.exists <- function( fname ) length(Sys.glob(fname))>0
  absolute.path.to.font.files <-  "/Users/bkotrc/font/";
  bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
```

Code F.2: MorphospaceFunctions.R (continued)

```
  ## if you do not have the correct font types
  for (i in 1:length(bera.names)) {
    stopifnot( file.exists(paste(absolute.path.to.font.files, bera.names[i], ".afm", sep="")) )
    stopifnot( file.exists(paste(absolute.path.to.font.files, bera.names[i], ".otf", sep="")) )
  }
  gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files, bera.names, ".afm", sep=""))
  pdfname <- "dq80100comparison.pdf"
  #Open PDF device
  pdf(file=pdfname, bg="white", width=(22.2/cm(1)), height=(14.8/cm(1)), pointsize=10, family=gillsans, colormodel="cmyk");
  axthck <- 0.3;
  layout(matrix(c(1,2,3,4), 2, 2, byrow = TRUE),widths=c(1.5,1))
  par(mar=c(4.1,4.1,1.5,4.1),oma=c(9.9,7.5,0,7.5), lwd=axthck, font.main=1, cex.main=1.5);
  #Define colors for 80% (pcol) and 100% (ncol) in plots
  pcol <- "royalblue"
  ncol <- "firebrick"
  #Plot chull results
  if(samplingmode == "in-bin" | samplingmode == "range-through")
  {
  #Set up axes, plot no points
plot(names(mpwd),res[1,],bty="n",type="n",xlim=c(65,0),ylim=c(-0.15,1),ylab="Normalized convex hull volume",xlab="Geologic Time (Ma)",axes=FALSE);
  axis(2, lwd=axthck);
  axis(1, lwd=axthck);
  title(main="A",adj=0.05,line=-1);
  #Results for 3D, filled circles
  points(names(mpwd),res[1,],pch=16,type="o",cex=0.8,lwd=1,col=pcol);
  points(names(mpwd),res2[1,],pch=15,type="o",cex=0.8,lwd=1,col=ncol);
  berg95(line=axthck);
legend(x=25,y=0.3,legend=c("80% threshold","100% threshold"),col=c(pcol,ncol),pch=c(16,15),box.lwd=axthck,lty="solid",pt.cex=.8,cex=0.75)
  }
  #Plot comparison of the two (colors for xlab, ylab)!
  plot(res[1,],res2[1,],type="n",xlab="",ylab="",bty="n",axes=FALSE,xaxs = "i",yaxs="i",xlim=c(0,1),ylim=c(0,1))
  title(main="B",adj=0.05,line=-1);
  mtext("Volume under 80% threshold",side=1,cex=0.8,line=3,col=pcol)
  mtext("Volume under 100% threshold",side=2,cex=0.8,line=3,col=ncol)
  axis(2, lwd=axthck);
  axis(1, lwd=axthck);
  #points(x=c(0,1),y=c(0,1),type="l",lty=2,lwd=axthck)
  points(res[1,],res2[1,],pch=3,lwd=0.8,xpd=TRUE)
  #Linear model (regression)
  results <- data.frame(res[1,],res2[1,])
  names(results) <- c("pco","nmds")
  model <- lm(nmds ~ pco,data = results)
  abline(model,lty=2,lwd=axthck)
  text(0.8,0.2,labels=bquote(R^2 == .(round(summary(model)$r.squared,digits=2))),font=3)
  #Plot MPWD results
  #Set up axes, plot no points
plot(names(mpwd),mpwd,bty="n",axes=FALSE,type="n",xlim=c(65,0),ylim=c(0.1,0.3),ylab="Mean pairwise distance",xlab="Geologic Time (Ma)");
  axis(2, lwd=axthck);
  axis(1, lwd=axthck);
  title(main="C",adj=0.05,line=-1);
```

Code F.2: MorphospaceFunctions.R (continued)

```
  #Results for (80%)
  points(names(mpwd),mpwd,pch=16,type="o",cex=0.8,lwd=1,col=pcol);
  #Results for (100%)
  points(names(mpwd),mpwd2,pch=15,type="o",cex=0.8,lwd=1,col=ncol);
  #Timescale
  berg95(line=axthck);
  #X-axis label
  #mtext("Geologic Time (Ma)",side=1,cex=0.8,line=3)
  legend(x=20,y=0.003,legend=c("80% threshold","100% threshold"),col=c(pcol,ncol),pch=c
       (16,15),box.lwd=axthck,lty="solid",pt.cex=.8,cex=0.75)
  #Plot comparison of the two (colors for xlab, ylab)!
  plot(mpwd,mpwd2,type="n",xlab="",ylab="",bty="n",axes=FALSE,xaxs = "i",yaxs="i",xlim=c
       (0.2,0.3),ylim=c(0.1,0.2))
  axis(2, lwd=axthck);
  axis(1, lwd=axthck);
  title(main="D",adj=0.05,line=-1);
  mtext("MPWD under 80% threshold",side=1,cex=0.8,line=3,col=pcol)
  mtext("MPWD under 100% threshold",side=2,cex=0.8,line=3,col=ncol)
  points(mpwd,mpwd2,pch=3,lwd=0.8,xpd=TRUE)
  #Linear model (regression)
  results2 <- data.frame(mpwd,mpwd2)
  names(results2) <- c("pco","nmds")
  model2 <- lm(nmds ~ pco,data = results2)
  abline(model2,lty=2,lwd=axthck)
  text(0.28,0.12,font=3,labels=bquote(R^2 == .(round(summary(model2)$r.squared,digits=2))))
  #Close graphics device
  dev.off();
  #Now embed font in that file
  embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font");
}
```

Code F.3: DiversityFunctions.R

```
#Functions used in the reanalysis of Rabosky & Sorhannus,
#2009 (Nature), for the subsampling of diatom diversity data
#from the Neptune database. Rewritten from functions received from
#D. Rabosky by email on 10/21/2009, keeping code mostly the
#same but renaming variables and adding comments, except SQS (new).
#Ben Kotrc, 11/5/2009 kotrc@fas.harvard.edu

#####

#Function to compute number of occurr. and lists per time bin,
#given the Neptune data (N), number of time bins (bins), and age
#range (agemax, agemin). Returns data frame rfinal, which
#contains columns:
#begin - lower bound on time bin
#end - upper bound on time bin
#nrecords - number of occurrences in time bin
#holes - number of holes in time bin
#lists - number of lists in time bin
#sqocc - sum of the number of occurrences, squared, in each list
#    in the time bin
#list.min - the length of the shortest list in the time bin
#lists.at/.pa - number of lists in the Atlantic/Pacific in bin
#occ.at/.pa - number of occurrences in Atlantic/Pacific in bin
#ints - number of time bins (same in each row!)
```

Code F.2: MorphospaceFunctions.R (continued)

```
#agemax - lower bound of oldest time bin (same in each row!)
#agemin - upper bound of youngest time bin (same in each row!)
#midpoint - mid-point age of each time bin
getSamplingIntensity <- function(N, bins, agemax, agemin)
{
  #Make an empty list called res
  res <- list();
  #Make an empty list called rfinal, this will contain results
  #returned by the function
  rfinal <- list();
  #Turn the list rfinal into a data frame
  rfinal <- as.data.frame(rfinal)
  #Set n_ints to bins, the number of time intervals
  n_ints <- bins;
  #Mysterious-set age limits to themselves?
  agemax <- agemax;
  agemin <- agemin;
  #Set the width of the time bin
  intsize <- (agemax-agemin)/n_ints;
  #Loop to calculate sampling intensity for each time bin, i
  for (i in 1:n_ints)
  {
    #Make a variable temp undefined
    temp <- NULL;
    #Empty the res variable
    res <- list();
    #Set the beginning age for the ith bin
    res$begin <- agemax - (i-1)*intsize;
    #Set the ending age for the ith bin
    res$end <- agemax - i*intsize;
    #Copy those occurrences that fall in the current (ith)
    #time bin from N to temp
    temp <- N[N$Sample.Age < res$begin & N$Sample.Age >= res$end, ];
    #How many occurrences in the ith time bin?
    res$nrecords <- nrow(temp);
    #How many sites in the ith time bin?
    res$holes <- length(unique(temp$Site));
    #How many lists in the ith time bin?
    res$lists <- length(unique(temp$List));
    #How many different species (diversity) in ith bin?
    res$div <- length(unique(temp$Species));
    #Make a table with the number of occurrences on each
    #list in the ith time bin
    ow <- table(temp$List);
    #Remove all columns of the table with 0 or fewer occurr.
    #(This doesn't seem necessary to me)
    ow <- ow[ow > 0];
    #Store the number of occurrences, squared, summed over
    #all lists in the ith time bin in res$sqocc
    res$sqocc <- sum(ow^2);
    #Store Good's u in res$goodsu
    res$goodsu <- goodsU(temp)

    #Calculate the smallest number of occurrences of any
    #list in the ith time bin
    #Are there any occurrences in the ith time bin?
    if (nrow(temp) > 0){
      #If so, store the smallest number of occurrences in
```

Code F.2: MorphospaceFunctions.R (continued)

```
      #any list in res$list.min
      res$list.min <- min(ow);
    }else{
      #If not, make res$list.min NA (missing value indic.)
      res$list.min <- NA;
    }

    #Get number of lists and occurrences by ocean
    #Number of lists in the Atlantic, in ith time bin
    res$lists.at <- length(unique(temp$List[temp$Ocean=='Atlantic']));
    #Number of lists in the Pacific, in ith time bin
    res$lists.pa <- length(unique(temp$List[temp$Ocean=='Pacific']));
    #Number of occurrences in the Atlantic
    res$occ.at <- length(temp$Ocean[temp$Ocean=='Atlantic']);
    #Number of occurrences in the Pacific
    res$occ.pa <- length(temp$Ocean[temp$Ocean=='Pacific']);

    #Tidy up results variables at end of for loop
    #Put the number of time bins into res$ints
    res$ints <- n_ints;
    #Put the bottom of the age bin into res$agemax
    res$agemax <- agemax;
    #Put the top of the age bin into res$agemin
    res$agemin <- agemin;
    #Convert the variable res to a data fram
    res <- as.data.frame(res);

    #Append the results for the ith time bin (res))
    rfinal <- rbind(rfinal, res);
    #End the for loop that goes through each time bin
  }
  #Calculate the midpoint of the time bins
  rfinal$midpoint <- (rfinal$begin + rfinal$end)/2;
  #Label the in-bin diversity values using the midpoint bin ages
  names(rfinal$div) <- rfinal$midpoint;
  #Return the results!
  return(rfinal)
}

#####

#Function to run a (single) subsampling of the full dataset using
#the algorithm of Foote & Miller (1996). This algorithm
#subsamples the full dataset by occurrences until a given quota of
#occurrences is reached. The function takes parameters:
#data - the full Neptune dataset to be subsampled
#intensity - the dataframe returned by getSamplingIntensity
#threshold - an integer value determining the number of
#occurrences to which the full dataset will be subsampled.
#Time bins with fewer occurrences than the threshold are discarded
#Returns:
#res - a pseudo-database with each time bin containing only
#    the threshold number of occurrences, randomly chosen; sub-
#    threshold bins omitted.
naiveRarefactionSubsample <- function(data, intensity, threshold)
{
  #Pre-allocate res, to hold subsampled database
  res <- NULL;
```

Code F.2: MorphospaceFunctions.R (continued)

```
  #Cull entries in intensity that don't meet the threshold
  intensity <- intensity[intensity$nrecords >= threshold, ];
  #Loop through each time bin
  for (i in 1:nrow(intensity)){
    #Choose the rows in the bin, put in temp
    temp <- data[data$Sample.Age > intensity$end[i] & data$Sample.Age <= intensity$begin[i
        ], ];
    #Subsample from temp to threshold number
    keepset <- temp[sample(1:nrow(temp),threshold),];
    #If this is the first time bin
    if (is.null(res)){
      #Put the subsampled rows into the output variable
      res <- keepset;
    }else{
      #If this is not the first time bin
      #Append the subsampled rows to the output database
      res <- rbind(res, keepset);
      }
    }
  #Return the results!
  return(res);
}

#####

#Function to run a (single) subsampling of the full dataset using
#the by-list, unweighted algorithm as described in Bush, Markey,
#and Marshall (2004). This algorithm subsamples the full dataset
#by lists, until a given quota of lists is reached (regardless of
#the number of occurrences in the resulting subsample).
#The function takes parameters:
#data - the full Neptune dataset to be subsampled
#intensity - the dataframe returned by getSamplingIntensity
#threshold - an integer value determining the number of
#lists to which the full dataset will be subsampled.
#Time bins with fewer occurrences than the threshold are discarded.
#Returns:
#res - a pseudo-database with each time bin containing only
#    occurrences drawn to a threshold number of lists, randomly
#    chosen; sub-threshold bins omitted.

byListUWSubsample <- function(data, intensity, threshold)
{
  #Pre-allocate res, to hold subsampled database
  res <- NULL;
  #Cull entries in intensity dataframe with insufficient lists
  intensity <- intensity[intensity$lists >= threshold, ];
  #Loop through each time bin
  for (i in 1:nrow(intensity)){
    #Take the relevant rows of the focal bin, put in temp
    temp <- data[data$Sample.Age > intensity$end[i] & data$Sample.Age <= intensity$begin[i
        ], ];
    #Subsample threshold number of lists from temp
    keepset <- temp[temp$List %in% sample(unique(temp$List), threshold), ];
    #If this is the first time bin
    if (is.null(res)){
      #Put the subsampled rows into the output variable
      res <- keepset;
```

Code F.2: MorphospaceFunctions.R (continued)

```
    }else{
      #If this is not the first time bin
      #Append the subsampled rows to the output database
      res <- rbind(res, keepset);
      }
    }
  #Return the results!
  return(res);
}

######

#Function to run a (single) subsampling of the full dataset using
#the by-list, weighted by occurrences-squared algorithm as
#described in Bush, Markey, and Marshall (2004). This algorithm
#subsamples the full dataset by lists, until a given quota of
#occurrences-squared is reached.
#The function takes parameters:
#data - the full Neptune dataset to be subsampled
#intensity - the dataframe returned by getSamplingIntensity
#threshold - an integer value determining the threshold number
#(= occurrences ^ 2) to which the full dataset will be subsampled
#Returns:
#res - a pseudo-database with each time bin containing only
#    occurrences drawn to a threshold number of occurrences
#    squared.
o2wSubsample <- function(data, intensity, threshold)
{
  #Initialize the return variable
  res <- NULL;
  #Cull the intensity dataframe to exclude bins with fewer
  #than the threshold number of occurrences^2
  intensity <- intensity[intensity$sqocc >= threshold, ];
  #Go through each row of the intensity dataframe,
  #subsample the corresponding time bin in the full dataset
  #to the threshold number of occurrences^2
  for (i in 1:nrow(intensity))
  {
    #Copy the rows of the full dataset in the
    #focal time bin into temp
    temp <- data[data$Sample.Age > intensity$end[i] & data$Sample.Age <= intensity$begin[i
        ], ];
    #Create a table x with the lists and list lengths
    #in the focal time bin
    x <- table(temp$List)[table(temp$List) > 0];
    #Make a data frame, list.attributes, with columns:
    #name (name of list), count (number of occurrences),
    #sqocc (number of occurrences, squared)
    list.attributes <- data.frame(name=names(x), count=x, sqocc=x^2, row.names=1:length(x))
        ;
    #Make a data frame with the list names in the full data
    #set in time bin i, i.e. lists available for subsampling
    lists.avail <- list.attributes$name;
    #Make a variable, set, to hold the result of subsampling
    #for time bin i
    set <- NULL;
    #Initialize a variable to keep track of how many
    #occurrences-squared are held in set (currently empty)
```

Code F.2: MorphospaceFunctions.R (continued)

```
    count <- 0;
    #Add lists to set until count exceeds the
    #sampling threshold of occurrences-squared
    while (count < threshold)
    {
      #Pick one random list from lists.avail, add to set
      set <- c(as.vector(sample(lists.avail, 1)), as.vector(set));
      #Update the available lists (lists.avail) by
      #subtracting out the list just added to the
      #subsample set
      lists.avail <- setdiff(lists.avail, set);

      #Keep track of how many occurrences-squared
      #are represented by the subsampled set
      count <- sum(list.attributes$sqocc[list.attributes$name %in% set]);
    }#End while subsampling
    #Copy the subsampled rows of the full-dataset time bin,
    #temp, into a new data frame, keepset
    keepset <- temp[temp$List %in% set, ];
    #If this is the first time bin looked at
    if (is.null(res))
    {
      #Put keepset into return variable
      res <- keepset;
    }
    else
    {
      #Append the keepset to the return variable
      #(i.e. add rows of subsampled data)
      res <- rbind(res, keepset);
    }#End if-else
  }#End loop through time bins
  #Send results back!
  return(res);
}

######

#Function to run a (single) subsampling of the full dataset using
#the by-list, weighted by occurrences algorithm as
#described in Bush, Markey, and Marshall (2004). This algorithm
#subsamples the full dataset by lists, until a given quota of
#occurrences-squared is reached.
#The function takes parameters:
#data - the full Neptune dataset to be subsampled
#intensity - the dataframe returned by getSamplingIntensity
#threshold - an integer value determining the threshold number
#of occurrences to which the full dataset will be subsampled
#Returns:
#res - a pseudo-database with each time bin containing only
#    occurrences drawn to a threshold number of occurrences
owSubsample <- function(data, intensity, threshold)
{
  #Initialize the return variable
  res <- NULL;
  #Cull the intensity dataframe to exclude bins with fewer
  #than the threshold number of occurrences^2
  intensity <- intensity[intensity$nrecords >= threshold, ];
```

Code F.2: MorphospaceFunctions.R (continued)

```
  #Go through each row of the intensity dataframe,
  #subsample the corresponding time bin in the full dataset
  #to the threshold number of occurrences^2
  for (i in 1:nrow(intensity))
  {
    #Copy the rows of the full dataset in the
    #focal time bin into temp
    temp <- data[data$Sample.Age > intensity$end[i] & data$Sample.Age <= intensity$begin[i
        ], ];
    #Create a table x with the lists and list lengths
    #in the focal time bin
    x <- table(temp$List)[table(temp$List) > 0];
    #Make a data frame, list.attributes, with columns:
    #name (name of list), count (number of occurrences),
    #sqocc (number of occurrences, squared)
    list.attributes <- data.frame(name=names(x), count=x, sqocc=x^2, row.names=1:length(x))
        ;
    #Make a data frame with the list names in the full data
    #set in time bin i, i.e. lists available for subsampling
    lists.avail <- list.attributes$name;
    #Make a variable, set, to hold the result of subsampling
    #for time bin i
    set <- NULL;
    #Initialize a variable to keep track of how many
    #occurrences-squared are held in set (currently empty)
    count <- 0;
    #Add lists to set until count exceeds the
    #sampling threshold of occurrences-squared
    while (count < threshold)
    {
      #Pick one random list from lists.avail, add to set
      set <- c(as.vector(sample(lists.avail, 1)), as.vector(set));
      #Update the available lists (lists.avail) by
      #subtracting out the list just added to the
      #subsample set
      lists.avail <- setdiff(lists.avail, set);

      #Keep track of how many occurrences
      #are represented by the subsampled set
      count <- sum(list.attributes$count[list.attributes$name %in% set]);
    }#End while subsampling
    #Copy the subsampled rows of the full-dataset time bin,
    #temp, into a new data frame, keepset
    keepset <- temp[temp$List %in% set, ];
    #If this is the first time bin looked at
    if (is.null(res))
    {
      #Put keepset into return variable
      res <- keepset;
    }
    else
    {
      #Append the keepset to the return variable
      #(i.e. add rows of subsampled data)
      res <- rbind(res, keepset);
    }#End if-else
  }#End loop through time bins
  #Send results back!
```

Code F.2: MorphospaceFunctions.R (continued)

```
  return(res);
}


######

#Function to run a (single) subsampling of the full dataset using
#the Shareholder Quorum Subsampling algorithm as described by Alroy
#2010 (Science, Palaeontology, and the Paleo Society Short Course
#volume), but with a modification that calculates Good's U using
#single-hole taxa rather than single-occurrence taxa (see goodsU()
#function in this file) to a threshold of coverage given by q.
#The function takes parameters:
#data - the full Neptune dataset to be subsampled
#intensity - the dataframe returned by getSamplingIntensity
#q - a value between zero and 1 representing the "quorum"
#    coverage to which the full dataset will be subsampled
#Returns:
#res - a pseudo-database with each time bin containing only
#    occurrences drawn to a threshold number of occurrences
sqsSubsample <- function(data, intensity, q)
{
  #Initialize the return variable
  res <- NULL;
  #Cull the intensity dataframe
  #Throw out bins with no occurrences
  intensity[intensity$nrecords == 0,] <- NA;
  intensity <- intensity[!is.na(intensity$nrecords),];
  #And those with less than the threshold coverage
  intensity[intensity$goodsu < q, ] <- NA;
  intensity <- intensity[!is.na(intensity$nrecords),];
  #Go through each row of the intensity dataframe,
  #subsample the corresponding time bin in the full dataset
  #to the threshold quorum
  for (i in 1:nrow(intensity))
  {
    #Copy the rows of the full dataset in the
    #focal time bin into temp
    temp <- data[data$Sample.Age > intensity$end[i] & data$Sample.Age <= intensity$begin[i
         ], ];

    #Calculate the target coverage (sum of frequencies) for this bin
    targetshares <- q/intensity$goodsu[i];

    #Total number of occurrences for this bin
    O <- intensity$nrecords[i];

    #Species list in this bin
    splist <- table(temp$Species);
    #Calculate their raw frequencies
    splistfreq <- splist/sum(splist);

    #Make a variable, set, to hold the rownames of occurrences subsampled
    #for time bin i
    set <- NULL;
    #Initialize a variable to keep track of sum of shares
    #of species in the subsample
    sharecount <- 0;
```

Code F.2: MorphospaceFunctions.R (continued)

```
    #Keep track of occurrences available for subsampling using their rownames()
    occs.avail <- rownames(temp);
    #Add occurrences to set until the sum of frequencies of the species,
    #countfreq, in the list exceeds the threshold coverage, target
    while (sharecount < targetshares)
    {
      #Pick one random occurrence from occs.avail, add to set
      set <- c(as.vector(sample(occs.avail, 1)), as.vector(set));
      #Update the available occurrences by
      #subtracting out the list just added to the
      #subsample set
      occs.avail <- setdiff(occs.avail, set);

      #Keep track of the sum of shares represented by the subsampled set
      #so far
      sharecount <- sum(splistfreq[unique(temp[set,"Species"])]);
    }#End while subsampling
    #Copy the subsampled rows of the full-dataset time bin,
    #temp, into a new data frame, keepset
    keepset <- temp[set, ];
    #If this is the first time bin looked at
    if (is.null(res))
    {
      #Put keepset into return variable
      res <- keepset;
    }
    else
    {
      #Append the keepset to the return variable
      #(i.e. add rows of subsampled data)
      res <- rbind(res, keepset);
    }#End if-else

  }#End loop through time bins
  #Send results back!
  return(res);
}


######

#Function to count number of 1-timers, 2-timers, etc. in a
#particular subsampled Neptune dataset
#Input arguments:
#data - the subsampled database, as obtained by
#   naiveRarefactionSubsampling.
#intensity - the result of getSamplingIntensity performed on the
#      full, unsubsampled dataset
#threshold - the number of occurrences to which the database was
#      subsampled.
#Returns res, a dataframe with columns:
#midpoint - time bin mid-point age
#begin - lower age bound on time bin
#end - upper age bound on time bin
#lists - number of lists in time bin
#ns - number of species in time bin (diversity)
#cores - number of cores in time bin
#nrecords - number of occurrences in time bin
```

Code F.2: MorphospaceFunctions.R (continued)

```
getStats.nr <-function(data, intensity, threshold)
{
  #Preallocate results variable
  res <- list();
  #Loop through each time bin, get bin diversities
  for (i in 1:nrow(intensity)){
    #Copy bin bounds and midpoint age to results variable
    res$midpoint[i] <- intensity$midpoint[i];
    res$begin[i] <- intensity$begin[i];
    res$end[i] <- intensity$end[i];
    #If there are more occurrences in the time bin than
    #the threshold
    if (intensity$nrecords[i] >= threshold){
      #Copy occurrences in bin from subsampled data to temp
      temp <- data[data$Sample.Age > intensity$end[i] & data$Sample.Age <= intensity$begin[
          i], ];
      #Count number of lists in time bin
      res$lists[i] <- length(unique(temp$List));      #Count number of unique species in
          temp (i.e. in bin)
      res$ns[i] <- length(unique(temp$Species));
      #Count number of unique holes in time bin
      res$cores[i] <- length(unique(temp$Hole.ID));
      #Count number of occurrences (surely, should=thresh?)
      res$nrecords[i] <- length(temp$Species);
    }#End if
    #If the time bin i is not quorate (insuff. occ.)
    else{
      #Make the entries of the results variable NA
      res$lists[i] <- NA;
      res$ns[i] <- NA;
      res$cores[i] <- NA;
      res$nrecords <- NA;
    }#End else
  }#End for loop

  #Set the first time bin (i=1) rate parameters (1-timers, etc)
  #to NA
  res <- setTimersNA(res, 1);

  #Loop through the time bins except for the first and last, and
  #count 1-timers, 2-timers, etc.
  for (i in 2:(nrow(intensity)-1)){
    #If previous, current, and next time bin are quorate
    if (intensity$nrecords[i-1] >= threshold & intensity$nrecords[i] >= threshold &
        intensity$nrecords[i+1] >= threshold){
      #Make dataframes containing species lists for bins
      #before, current, and after bin i
      #List of species in bin i-1
      prebin <- unique(data$Species[data$Sample.Age > intensity$end[i-1] & data$Sample.Age
          <= intensity$begin[i-1]]);
      #List of species in bin i
      inbin <- prebin <- unique(data$Species[data$Sample.Age > intensity$end[i] & data$
          Sample.Age <= intensity$begin[i]]);
      #List of species in bin i+1
      postbin <- unique(data$Species[data$Sample.Age > intensity$end[i+1] & data$Sample.Age
          <= intensity$begin[i+1]]);

      #Now count x-timers for bin i
```

Code F.2: MorphospaceFunctions.R (continued)

```
      #1-timers
      res$t1[i] <- length(setdiff(setdiff(inbin, prebin),postbin));
      #2-timers, in previous bin
      res$t2ib[i] <- length(prebin[prebin %in% inbin]);
      #2-timers, in following bin
      res$t2ia[i] <- length(inbin[inbin %in% postbin]);
      #3-timers
      res$t3[i] <- length(prebin[prebin %in% postbin[postbin %in% inbin]]);
      #Part-timers, (before and after, but not in bin)
      res$tp[i] <- length(setdiff(prebin[prebin %in% postbin], inbin));
      }#End quorate if
      #If time bin i is not quorate
      else{
        #Set timers to NA values
        res <- setTimersNA(res, i);
      }#End else
    }#End for loop through time bins

  #Finally, set values to NA for last time bin
  res <- setTimersNA(res, nrow(intensity));

  #Return results!
  return(as.data.frame(res));
}

######

#Function to count number of 1-timers, 2-timers, etc. in a
#particular subsampled Neptune dataset
#Input arguments:
#data - the subsampled database, as obtained by
#   byListUWSubsample.
#intensity - the result of getSamplingIntensity performed on the
#      full, unsubsampled dataset
#threshold - the number of lists to which the database was
#      subsampled.
#Returns res, a dataframe with columns:
#midpoint - time bin mid-point age
#begin - lower age bound on time bin
#end - upper age bound on time bin
#lists - number of lists in time bin
#ns - number of species in time bin (diversity)
#cores - number of cores in time bin
#nrecords - number of occurrences in time bin
getStats.uw <-function(data, intensity, threshold)
{
  #Preallocate results variable
  res <- list();
  #Loop through each time bin, get bin diversities
  for (i in 1:nrow(intensity)){
    #Copy bin bounds and midpoint age to results variable
    res$midpoint[i] <- intensity$midpoint[i];
    res$begin[i] <- intensity$begin[i];
    res$end[i] <- intensity$end[i];
    #If there are more lists in the unsubsampled data set in
    #the focal time bin than the threshold
    if (intensity$lists[i] >= threshold){
      #Copy occurrences in bin from subsampled data to temp
```

Code F.2: MorphospaceFunctions.R (continued)

```
592       temp <- data[data$Sample.Age > intensity$end[i] & data$Sample.Age <= intensity$begin[
              i], ];
593       #Count number of lists in time bin
594       res$lists[i] <- length(unique(temp$List));      #Count number of unique species in
              temp (i.e. in bin)
595       res$ns[i] <- length(unique(temp$Species));
596       #Count number of unique holes in time bin
597       res$cores[i] <- length(unique(temp$Hole.ID));
598       #Count number of occurrences (surely, should=thresh?)
599       res$nrecords[i] <- length(temp$Species);
600     }#End if
601     #If the time bin i is not quorate (insuff. lists)
602     else{
603       #Make the entries of the results variable NA
604       res$lists[i] <- NA;
605       res$ns[i] <- NA;
606       res$cores[i] <- NA;
607       res$nrecords <- NA;
608     }#End else
609   }#End for loop
610
611   #Set the first time bin (i=1) rate parameters (1-timers, etc)
612   #to NA
613   res <- setTimersNA(res, 1);
614
615   #Loop through the time bins except for the first and last, and
616   #count 1-timers, 2-timers, etc.
617   for (i in 2:(nrow(intensity)-1)){
618     #If previous, current, and next time bin are quorate
619     if (intensity$lists[i-1] >= threshold & intensity$lists[i] >= threshold & intensity$
          lists[i+1] >= threshold){
620       #Make dataframes containing species lists for bins
621       #before, current, and after bin i
622       #List of species in bin i-1
623       prebin <- unique(data$Species[data$Sample.Age > intensity$end[i-1] & data$Sample.Age
              <= intensity$begin[i-1]]);
624       #List of species in bin i
625       inbin <- prebin <- unique(data$Species[data$Sample.Age > intensity$end[i] & data$
              Sample.Age <= intensity$begin[i]]);
626       #List of species in bin i+1
627       postbin <- unique(data$Species[data$Sample.Age > intensity$end[i+1] & data$Sample.Age
               <= intensity$begin[i+1]]);
628
629       #Now count x-timers for bin i
630       #1-timers
631       res$t1[i] <- length(setdiff(setdiff(inbin, prebin),postbin));
632       #2-timers, in previous bin
633       res$t2ib[i] <- length(prebin[prebin %in% inbin]);
634       #2-timers, in following bin
635       res$t2ia[i] <- length(inbin[inbin %in% postbin]);
636       #3-timers
637       res$t3[i] <- length(prebin[prebin %in% postbin[postbin %in% inbin]]);
638       #Part-timers, (before and after, but not in bin)
639       res$tp[i] <- length(setdiff(prebin[prebin %in% postbin], inbin));
640       }#End quorate if
641       #If time bin i is not quorate
642       else{
643         #Set timers to NA values
```

Code F.2: MorphospaceFunctions.R (continued)

```
644         res <- setTimersNA(res, i);
645       }#End else
646     }#End for loop through time bins
647 
648   #Finally, set values to NA for last time bin
649   res <- setTimersNA(res, nrow(intensity));
650 
651   #Return results!
652   return(as.data.frame(res));
653 }
654 
655 ######
656 
657 #Function to count number of 1-timers, 2-timers, etc. in a
658 #particular subsampled Neptune dataset
659 #Input arguments:
660 #data - the subsampled database, as obtained by
661 #   o2wSubsample.
662 #intensity - the result of getSamplingIntensity performed on the
663 #      full, unsubsampled dataset
664 #threshold - the number of occurrences-squared to which the
665 #      database was subsampled.
666 #Returns res, a dataframe with columns:
667 #midpoint - time bin mid-point age
668 #begin - lower age bound on time bin
669 #end - upper age bound on time bin
670 #lists - number of lists in time bin
671 #ns - number of species in time bin (diversity)
672 #cores - number of cores in time bin
673 #nrecords - number of occurrences in time bin
674 getStats.o2w <-function(data, intensity, threshold)
675 {
676   #Preallocate results variable
677   res <- list();
678   #Loop through each time bin, get bin diversities
679   for (i in 1:nrow(intensity)){
680     #Copy bin bounds and midpoint age to results variable
681     res$midpoint[i] <- intensity$midpoint[i];
682     res$begin[i] <- intensity$begin[i];
683     res$end[i] <- intensity$end[i];
684     #If there are more occ-sq in the unsubsampled data set in
685     #the focal time bin than the threshold
686     if (intensity$sqocc[i] >= threshold){
687       #Copy occurrences in bin from subsampled data to temp
688       temp <- data[data$Sample.Age > intensity$end[i] & data$Sample.Age <= intensity$begin[
              i], ];
689       #Count number of lists in time bin
690       res$lists[i] <- length(unique(temp$List));      #Count number of unique species in
              temp (i.e. in bin)
691       res$ns[i] <- length(unique(temp$Species));
692       #Count number of unique holes in time bin
693       res$cores[i] <- length(unique(temp$Hole.ID));
694       #Count number of occurrences (surely, should=thresh?)
695       res$nrecords[i] <- length(temp$Species);
696     }#End if
697     #If the time bin i is not quorate (insuff. lists)
698     else{
699       #Make the entries of the results variable NA
```

Code F.2: MorphospaceFunctions.R (continued)

```
      res$lists[i] <- NA;
      res$ns[i] <- NA;
      res$cores[i] <- NA;
      res$nrecords <- NA;
    }#End else
  }#End for loop

  #Set the first time bin (i=1) rate parameters (1-timers, etc)
  #to NA
  res <- setTimersNA(res, 1);

  #Loop through the time bins except for the first and last, and
  #count 1-timers, 2-timers, etc.
  for (i in 2:(nrow(intensity)-1)){
    #If previous, current, and next time bin are quorate
    if (intensity$sqocc[i-1] >= threshold & intensity$sqocc[i] >= threshold & intensity$
        sqocc[i+1] >= threshold){
      #Make dataframes containing species lists for bins
      #before, current, and after bin i
      #List of species in bin i-1
      prebin <- unique(data$Species[data$Sample.Age > intensity$end[i-1] & data$Sample.Age
          <= intensity$begin[i-1]]);
      #List of species in bin i
      inbin <- prebin <- unique(data$Species[data$Sample.Age > intensity$end[i] & data$
          Sample.Age <= intensity$begin[i]]);
      #List of species in bin i+1
      postbin <- unique(data$Species[data$Sample.Age > intensity$end[i+1] & data$Sample.Age
          <= intensity$begin[i+1]]);

      #Now count x-timers for bin i
      #1-timers
      res$t1[i] <- length(setdiff(setdiff(inbin, prebin),postbin));
      #2-timers, in previous bin
      res$t2ib[i] <- length(prebin[prebin %in% inbin]);
      #2-timers, in following bin
      res$t2ia[i] <- length(inbin[inbin %in% postbin]);
      #3-timers
      res$t3[i] <- length(prebin[prebin %in% postbin[postbin %in% inbin]]);
      #Part-timers, (before and after, but not in bin)
      res$tp[i] <- length(setdiff(prebin[prebin %in% postbin], inbin));
      }#End quorate if
      #If time bin i is not quorate
      else{
        #Set timers to NA values
        res <- setTimersNA(res, i);
      }#End else
    }#End for loop through time bins

  #Finally, set values to NA for last time bin
  res <- setTimersNA(res, nrow(intensity));

  #Return results!
  return(as.data.frame(res));
}

#####

#Function to set the 1-timers, 2-timers, 3-timers, part-timers
```

Code F.2: MorphospaceFunctions.R (continued)

```
#variables to "NA" value for time bins that do not have enough
#occurrences while running the getStats.nr function
#Takes arguments:
#res - the results dataframe returned by the getStats.nr function,
#    (see that function for components)
#index - the time bin for which to set values to NA
setTimersNA <- function(res, index)
{
  res$t1[index] <- NA;
  res$t2ib[index] <- NA;
  res$t2ia[index] <- NA;
  res$t3[index] <- NA;
  res$tp[index] <- NA;
  return(res);
}

#####

#Function to generate range-through diversity from a dataset
#Takes arguments:
#N - the raw Neptune database
#age.min - lower bound of sample ages to be examined
#age.max - upper bound of sample ages to be examined
#bins - number of age bins into which to divide samples for taxon enumeration (species
     counting)
#Returns:
#divrt - a vector representing diversities in time bins
rangeThrough  <- function(N, age.min, age.max, bins)
{
  #Preallocate results variable
  res <- list();
  #Calculate the size of the time bins
  binsize <- (age.max-age.min)/bins;
  #First and last bins have no range-through (obviously)
  focalbin <- N[N$Sample.Age <= binsize & N$Sample.Age > age.min,];
  res <- length(unique(focalbin$Species));
  #Loop through each time bin, get bin diversities
  for (i in 2:(bins-1)){
    #Calculate focal bin boundaries
    binstart <- i*binsize;
    binend <- (i-1)*binsize;
    #Divide data set into before, during, after bins
    beforebin <- N[N$Sample.Age > binstart,];
    focalbin <- N[N$Sample.Age <= binstart & N$Sample.Age > binend,];
    afterbin <- N[N$Sample.Age <= binend,];
    #Count taxa in focal bin
    inbin <- length(unique(focalbin$Species));
    #Count taxa before and after, but not in focal bin
    twotimers <- length(unique(setdiff(beforebin$Species[beforebin$Species %in% afterbin$
         Species],focalbin$Species)));
    #Add counts to get range-through diversity
    res <- c(res,inbin + twotimers);
  }
  #Add diversity for last bin
  focalbin <- N[N$Sample.Age <= age.max & N$Sample.Age > (bins-1)*binsize,];
  res[bins] <- length(unique(focalbin$Species));
  #Add bin ages as "names" attribute of res variable
  names(res) <- seq(from=age.min+(binsize/2),to=age.max-(binsize/2),by=binsize);
```

Code F.2: MorphospaceFunctions.R (continued)

```
  #Send back vector of results
  return(res);
}

#####

#Function to calculate "preservation" 3-timer statistic for time
#bins. What is the proportion of taxa found either side of the
#focal bin, but not in the focal bin itself? This is essentially
#the difference between in-bin and range-through diversity as a
#proportion of in-bin diversity (but not quite, because this
#method does not count 1- or 2-timers).
#Takes arguments:
#N - the raw Neptune database
#age.min - lower bound of sample ages to be examined
#age.max - upper bound of sample ages to be examined
#bins - number of age bins into which to divide samples for taxon enumeration (species counting)
#Returns:
#res - a data frame with the following columns (each row 1 t bin):
#res$res - the ratio of 3t/(3t+pt)
#res$res3t - the number of three timers in the time bin
#res$respt - the number of part timers in the time bin
preservationThreeTimers  <- function(N, age.min, age.max, bins)
{
  #Preallocate results variable
  res <- list();
  res3t <- list();
  respt <- list();
  #Calculate the size of the time bins
  binsize <- (age.max-age.min)/bins;
  #First and last bins have no range-through (obviously)
  focalbin <- N[N$Sample.Age <= binsize & N$Sample.Age > age.min,];
  res <- NA;
  res3t <- NA;
  respt <- NA;
  #Loop through each time bin, get bin diversities
  for (i in 2:(bins-1)){
    #Calculate focal bin boundaries
    binstart <- i*binsize;
    binend <- (i-1)*binsize;
    #Divide data set into before, during, after bins
    beforebin <- N[N$Sample.Age > binstart,];
    focalbin <- N[N$Sample.Age <= binstart & N$Sample.Age > binend,];
    afterbin <- N[N$Sample.Age <= binend,];
    #Count taxa in focal bin also present before and after
    threetimers <- length(beforebin$Species[beforebin$Species %in% afterbin$Species[afterbin$Species %in% focalbin$Species]]);
    res3t <- c(res3t,threetimers);
    #Count taxa before and after, but not in focal bin
    parttimers <- length(unique(setdiff(beforebin$Species[beforebin$Species %in% afterbin$Species],focalbin$Species)));
    respt <- c(respt,parttimers);
    #Add counts to three-timer statistic, if
    #counts aren't both zero
    if((threetimers) > 0){
      res <- c(res,(threetimers/(threetimers+parttimers)));
      }
```

Code F.2: MorphospaceFunctions.R (continued)

```
865       else{res <- c(res, NA)}
866   }
867   #Add NA for last bin
868   res[bins] <- NA;
869   res3t[bins] <- NA;
870   respt[bins] <- NA;
871   #Send back vector of results
872   result <- as.data.frame(res);
873   result$res3t <- res3t;
874   result$respt <- respt;
875   return(result);
876 }
877 
878 #####
879 
880 #Function to calculate Alroy's 3-timer statistic for time
881 #bins. What is the proportion of taxa found in the adjacent time
882 #bin on either side of the focal bin, but not in the focal bin
883 #itself?
884 #Takes arguments:
885 #N - the raw Neptune database
886 #age.min - lower bound of sample ages to be examined
887 #age.max - upper bound of sample ages to be examined
888 #bins - number of age bins into which to divide samples for taxon enumeration (species
        counting)
889 #Returns:
890 #res - a data frame with the following columns (each row 1 t bin):
891 #res$res - the ratio of 3t/(3t+pt)
892 #res$res3t - the number of three timers in the time bin
893 #res$respt - the number of part timers in the time bin
894 alroyThreeTimers  <- function(N, age.min, age.max, bins)
895 {
896   #Preallocate results variable
897   res <- list();
898   res3t <- list();
899   respt <- list();
900   #Calculate the size of the time bins
901   binsize <- (age.max-age.min)/bins;
902   #First and last bins have no range-through (obviously)
903   focalbin <- N[N$Sample.Age <= binsize & N$Sample.Age > age.min,];
904   res <- NA;
905   res3t <- NA;
906   respt <- NA;
907   #Loop through each time bin, get bin diversities
908   for (i in 2:(bins-1)){
909     #Calculate focal bin boundaries
910     binstart <- i*binsize;
911     binend <- (i-1)*binsize;
912     #Divide data set into before, during, after bins
913     beforebin <- N[N$Sample.Age > binstart & N$Sample.Age <= binstart+binsize,];
914     focalbin <- N[N$Sample.Age <= binstart & N$Sample.Age > binend,];
915     afterbin <- N[N$Sample.Age <= binend & N$Sample.Age > binend-binsize,];
916     #Count taxa in focal bin also present before and after
917     threetimers <- length(beforebin$Species[beforebin$Species %in% afterbin$Species[
          afterbin$Species %in% focalbin$Species]]);
918     res3t <- c(res3t,threetimers);
919     #Count taxa before and after, but not in focal bin
```

Code F.2: MorphospaceFunctions.R (continued)

```
    parttimers <- length(unique(setdiff(beforebin$Species[beforebin$Species %in% afterbin$
        Species],focalbin$Species)));
    respt <- c(respt,parttimers);
    #Add counts to three-timer statistic, if
    #counts aren't both zero
    if((threetimers) > 0){
      res <- c(res,(threetimers/(threetimers+parttimers)));
      }
      else{res <- c(res, NA)}
  }
  #Add NA for last bin
  res[bins] <- NA;
  res3t[bins] <- NA;
  respt[bins] <- NA;
  #Send back vector of results
  result <- as.data.frame(res);
  result$res3t <- res3t;
  result$respt <- respt;
  return(result);
}

#####

#Function to calculate Good's u (Good, 1953), an estimator of the
#taxonomic diversity "coverage" of a sample.
#Takes argument:
#data - a Neptune database or subsample thereof
#Returns:
#u - Good's u estimator
goodsU <- function(data)
{
  #How many occurrences are in the sample in total
  O <- length(data$Species);

  #How many species are in the sample only once? i.e. o1
  #o1 <- length(table(data$Species)[table(data$Species)==1]);

  #Find number of single-hole species, h1
  #Make a list of all the species
  splist <- unique(data$Species);
  #Set the number of single hole taxa to zero
  h1 <- 0;
  #Loop through each species and find out if it's from one hole only
  for(i in 1:length(splist))
  {
    #Is this taxon in more than one hole?
    if(length(unique(data[data$Species == splist[i],]$Hole.ID)) == 1)
    {
      #Count it as a single-hole taxon
      h1 <- h1+1;
    }
  }

  #Calculate Good's u, using the single-hole taxon modification
  u <- 1-(h1/O);

  #Send it back!
  return(u);
```

Code F.2: MorphospaceFunctions.R (continued)

```
977 }
```

G

R Code for RadData Database

G.1 R script for creating RadData database

Code G.1: CreateRadData.R

```
#Code to create a relational database for managing radiolarian #morphometric data. The
     schema for this database is contained in a
#separate file named RadData Schema.pdf.
#Benjamin Kotrc, February 2011, kotrc@fas.harvard.edu

#Load the necessary libraries for using SQLite with R
library(DBI);
library(RSQLite);

#Change working directory to point to the directory where RadData
#database file is stored.
setwd('/Users/Ben/Dropbox/Harvard/By-Lineage\ Rads/RadData\ Database');

#Create a new connection with a database (if the file doesn't
#exist yet, it'll be created). Returns handle for database
#connection.
con <- dbConnect(SQLite(), "RadData.sqlite");

#Create the first table, Holes
dbGetQuery(con, "CREATE TABLE holes (hole_id CHAR(10) PRIMARY KEY, latitude REAL, longitude
      REAL)");

#Now the next table, Slides
dbGetQuery(con, "CREATE TABLE slides (slide_id CHAR(16), hole_id CHAR(10) REFERENCES holes,
      depth REAL, age REAL, preservation CHAR(2), PRIMARY KEY(slide_id))");
```

Code G.1: CreateRadData.R (continued)

```
23
24 #Now the penultimate table, Individuals
25 dbGetQuery(con, "CREATE TABLE individuals (indiv_id INTEGER PRIMARY KEY, slide_id CHAR(16)
       REFERENCES slides, species TEXT, comment TEXT)");
26
27
28 #Now let's set up the final table, Measurements
29 dbGetQuery(con, "CREATE TABLE measurements (meas_id INTEGER PRIMARY KEY, indiv_id INTEGER
       REFERENCES individuals, meas_type TEXT, meas_value REAL, imagefile TEXT)");
30
31 #Make sure to close the connection before you quit.
32 dbDisconnect(con);
```

G.2 R script for interface to RadData database

Code G.2: RadDataInterface.R

```
1  ######################################################################
2  #Benjamin Kotrc, February 2011, kotrc@fas.harvard.edu
3  ######################################################################
4  #Interface to collect radiolarian morphometric data for storage
5  #in the RadData database, using the Canon EOS Utility and ImageJ,
6  #with associated macros. A schematic sketch of the workflow can be
7  #found in the file Interface Design.pdf. This script requires the
8  #RadData.sqlite file to have been set up using the CreateRadData.R
9  #script. The design of the RadData database can be found in the
10 #RadData Schema.pdf file.
11 ######################################################################
12
13
14
15 ######################################################################
16 #Functions:
17 ######################################################################
18
19 ######################################################################
20 #Function to take a row of ImageJ measurement data, as taken from
21 #the "pipe" file, with the associated measurement type ID, and
22 #turn it into a string that can be passed to SQL to execute an
23 #INSERT to create a row for the measurement in the 'measurements'
24 #table of the RadData database.
25 #Takes arguments:
26 #data     - A data frame with a single row, containing
27 #         $V1 - the measurement value
28 #         $V2 - the name of the associated image file
29 #         $measOrder - the ID of the measurement type
30 #current_indiv_ID - the primary key field of the individual with
31 #         which these data are associated
32 #Returns:
33 #sql    - A string of SQL that will make an insertion into the
34 #       RadData database's 'measurements' table
35 createSQLInsert <- function(data,current_indiv_ID)
36 {
```

Code G.2: RadDataInterface.R (continued)

```
  sql <- paste("INSERT INTO measurements VALUES (NULL,",current_indiv_ID,",",data$measOrder
       ,",",data$V1,",'",data$V2,"')",sep="");
  return(sql);
}

#####################################################################
#Function to look up the string describing the measurement type
#for a given measurement type identifier
#Takes arguments:
#measNum - An integer between 1 and 52, ID of a measurement type
#Returns:
#measName - A string describing the measurement type
measName <- function(measNum)
{
  names <- c("Width of spongy column, distal",
"Width of spongy column, proximal",
"Width of spongy columns",
"Width of cortical shell",
"Width of cortical shell protrusions",
"Width of medullary shell",
"Width of inner medullary shell",
"Width of outer medullary shell",
"Width of polar caps",
"Width of veil",
"Width of horn",
"Width of cephalis",
"Width of thorax",
"Width of thorax (at middle)",
"Width of abdominal stricture (external)",
"Width of abdomen",
"Width of base",
"Width of ante-/postcephalic chamber",
"Width of shell at base of eucephalic lobe",
"Length of cortical shell",
"Length of cortical shell protrusions",
"Length of spongy columns",
"Length of polar caps",
"Length of veil",
"Length of horn",
"Length from base of horn to widest part of thorax",
"Length from widest part of thorax to widest part of abdomen",
"Length from widest part of abdomen to base",
"Length from collar stricture to widest part of thorax",
"Length from collar stricture to widest part of abdomen",
"Length from widest part of abdomen to base",
"Length from widest part of thorax to abdominal stricture",
"Length from abdominal stricture to base",
"Length of whole shell",
"Length of thorax",
"Length from top of shell to base of ante-/postcephalic chamber",
"Length from top of shell to base of eucephalic lobe",
"Pore area on cortical shell",
"Pore area on spongy columns",
"Pore area on veil",
"Pore area on polar cap",
"Pore area on horn",
"Pore area on thorax",
"Pore area on abdomen",
```

Code G.2: RadDataInterface.R (continued)

```
"Pore area on post-abdominal segment(s)",
"Shell thickness of cortical shell",
"Shell thickness of veil",
"Shell thickness of polar cap",
"Shell thickness of thorax",
"Shell thickness of abdomen",
"Shell thickness of post-abdominal segment(s)",
"Width of shell at top of shell",
"Length of tubercles or plicae",
"Number of tubercles or plicae",
"Number of polar caps",
"Shell thickness area (top view)",
"Length of cephalis",
"Length of abdomen");

  return(names[measNum]);
}

#####################################################################
#Function to look up the order in which measurements are expected
#for a given species
#Takes arguments:
#new_indiv_sp - the species name
#Returns:
#measOrder - A vector of integers representing the measurement
#      types for the species
getMeasOrder <- function(new_indiv_sp)
{
  #Define shorthands for the species name strings
  #Didymocyrtis-Diartus
  La <- 'Lithocyclia angusta';
  Dpr <- 'Didymocyrtis prismatica';
  Dv <- 'Didymocyrtis violina';
  Dl <- 'Didymocyrtis laticonus';
  Dm <- 'Didymocyrtis mammifera';
  Dan <- 'Didymocyrtis antepenultima';
  Dpen <- 'Didymocyrtis penultima';
  Dt <- 'Didymocyrtis tetrathalamus';
  Dav <- 'Didymocyrtis avita';
  Dpet <- 'Diartus petterssoni';
  Dh <- 'Diartus hughesi';
  #Artophormis
  Ab <- 'Artophormis barbadensis';
  Ag <- 'Artophormis gracilis';
  #Stichocorys
  Sd <- 'Stichocorys delmontensis';
  Sp <- 'Stichocorys peregrina';
  Sw <- 'Stichocorys wolffii';
  #Centrobotrys
  Cg <- 'Centrobotrys gravida';
  Cp <- 'Centrobotrys petrushevskayae';
  Ct <- 'Centrobotrys thermophila';
  #Phormocyrtis
  Pse <- 'Phormocyrtis striata exquisita';
  Pss <- 'Phormocyrtis striata striata';
  #Theocyrtis
  Tt <- 'Theocyrtis tuberosa';
  Ta <- 'Theocyrtis annosa';
```

Code G.2: RadDataInterface.R (continued)

```
152
153
154   if(new_indiv_sp == La){
155     measOrder <- c(1,2,4,6,22,38,39,46);
156   }else if(new_indiv_sp == Dpr){
157     measOrder <- c(1,2,4,6,22,20,39,38,46);
158   }else if(new_indiv_sp == Dv){
159     measOrder <- c(1,2,4,7,8,22,20,53,39,38,46);
160   }else if(new_indiv_sp == Dm){
161     measOrder <- c(1,2,4,7,8,22,20,53,39,38,46);
162   }else if(new_indiv_sp == Dan){
163     measOrder <- c(1,2,9,4,7,8,22,23,20,53,39,38,46,10,24,40,47);
164   }else if(new_indiv_sp == Dpen){
165     measOrder <- c(1,2,9,4,7,8,22,23,20,53,39,38,46,10,24,40,47);
166   }else if(new_indiv_sp == Dl){
167     measOrder <- c(1,2,9,4,7,8,22,23,20,53,39,38,46,10,24,40,47);
168   }else if(new_indiv_sp == Dt){
169     measOrder <- c(9,4,23,20,7,8,41,38,48,46,10,24,40,47);
170   }else if(new_indiv_sp == Dav){
171     measOrder <- c(1,2,9,4,7,8,22,23,20,53,39,38,46,10,24,40,47);
172   }else if(new_indiv_sp == Dpet){
173     measOrder <- c(3,5,4,7,8,22,21,20,39,38,46);
174   }else if(new_indiv_sp == Dh){
175     measOrder <- c(9,4,8,7,23,20,38,46);
176   }else if(new_indiv_sp == Ab){
177     measOrder <- c(25,11,42,12,13,15,29,32,33,43,44,45,49,50,51);
178   }else if(new_indiv_sp == Ag){
179     measOrder <- c(25,11,42,12,13,15,29,32,33,43,44,45,49,50,51);
180   }else if(new_indiv_sp == Sd){
181     measOrder <- c(25,11,12,16,17,30,28,44,45,50,51);
182   }else if(new_indiv_sp == Sp){
183     measOrder <- c(25,11,12,16,17,30,28,44,45,50,51);
184   }else if(new_indiv_sp == Sw){
185     measOrder <- c(25,11,12,16,17,30,28,44,45,50,51);
186   }else if(new_indiv_sp == Cg){
187     measOrder <- c(34,35,36,18,13,12,43,49);
188   }else if(new_indiv_sp == Cp){
189     measOrder <- c(34,35,36,18,14,12,43,49);
190   }else if(new_indiv_sp == Ct){
191     measOrder <- c(37,35,52,12,19,17,43,49);
192   }else if(new_indiv_sp == Pss | new_indiv_sp == Pse){
193     measOrder <- c(25,26,27,28,11,13,16,17,43,44,49,50);
194   }else if(new_indiv_sp == Tt | new_indiv_sp == Ta){
195     measOrder <- c(25,11,57,12,35,13,58,16,43,49,53,56);
196   }
197   return(measOrder);
198 }
199
200 ######################################################################
201 #Function to read a set of measurements made in ImageJ and written
202 #to the ”pipe” text file, and return them as a data frame.
203 #Takes arguments:
204 #measOrder - A vector containing the names of the measurements
205 #      expected in the file, in the expected order
206 #Returns:
207 #measurements - A data frame containing, for each measurement,
208 #       the value, associated file name, and measurement
209 #       name; measurements with zero value omitted
```

Code G.2: RadDataInterface.R (continued)

```
#lengthOK - A boolean variable indicating whether the expected
#     number of measurements were found in the "pipe" file
readMeasurements <- function(measOrder)
{
  #Read the pipe file into a data frame called meas
  meas <- read.delim("pipefilename.txt",header = FALSE,as.is=TRUE);
  #Make sure file is of right length
  if(nrow(meas)==length(measOrder)){
    lengthOK <- TRUE;
    #Do the thing
    measurements <- cbind(meas,measOrder);
    #Cull negative-valued measurements, which represent NAs (missing data)
    measurements <- measurements[measurements$V1 >= 0,];
  #If the file is not of the right length
  }else{
    #Raise the red flag
    lengthOK <- FALSE;
    measurements <- NA;
  }
  return(list(lengthOK,measurements));
}

######################################################################
#Function to exit out of menu, and close the database connection.
#Returns FALSE value to signal termination of while loop.
goodbye  <- function(con)
{
  #Close connection to the database
  dbDisconnect(con);
  #Polite goodbye to the user
  cat(rep("\n",17),"Thank you for your hard work. You are one step closer to graduating.\n\
      nCome back again soon.\n\n\n\n\n", sep="");
  #Close the graphics window
  dev.off();
  return(FALSE);
}

######################################################################
#Function to export the RadData database as a set of .CSV files.
saveBackup  <- function(con)
{
  #Get current date
  date <- format(Sys.time(), "%Y %b %d %H.%M");
  #Create a new directory in backups, named after current date
  dir.create(paste("Backups/",date,sep=""));
  #Write each table to file in the folder with current date
  write.table(dbGetQuery(con, "SELECT * FROM holes"),paste("Backups/",date,"/holes.txt",sep
      =""),row.names=FALSE,col.names=TRUE,sep="\t",quote=FALSE);
  write.table(dbGetQuery(con, "SELECT * FROM individuals"),paste("Backups/",date,"/
      individuals.txt",sep=""),row.names=FALSE,col.names=TRUE,sep="\t",quote=FALSE);
  write.table(dbGetQuery(con, "SELECT * FROM measurements"),paste("Backups/",date,"/
      measurements.txt",sep=""),row.names=FALSE,col.names=TRUE,sep="\t",quote=FALSE);
  write.table(dbGetQuery(con, "SELECT * FROM slides"),paste("Backups/",date,"/slides.txt",
      sep=""),row.names=FALSE,col.names=TRUE,sep="\t",quote=FALSE);
  print(paste("Backups written to tab-separated text files in the directory named '",date,"
      '."));
}

```

Code G.2: RadDataInterface.R (continued)

```
#####################################################################
#Function to set the slide to be worked on. Either creates an
#entry in the slides table of RadData, or returns an existing one.
#Returns slide_id value of this slide.
setSlide <- function(con)
{
  #Ask if this is a new slide
  cat("\nIs this a new slide?\n");
  switch(menu(c("Yes, new slide", "No, existing slide")) + 1, cat("Nothing done\n"), choice
       <- "new", choice <- "existing");
  if(choice=="existing")
  {
    print("These slides are in the database:");
    list <- dbGetQuery(con, "SELECT slide_id FROM slides")$slide_id;
    for(i in 1:length(list))
    {
      cat("\n",i,": ",list[i])
    }
    slidepick <- readline("Which number slide would you like to work on? ");
    new_slide_id <- dbGetQuery(con, "SELECT slide_id FROM slides")$slide_id[as.numeric(
        slidepick)];
  }
  if(choice=="new")
  {
    #Prompt for slide ID
    new_slide_id <- readline("Please enter the slide ID for this slide (e.g. 0709C
        ,27,1,103-110): ");
    #Check to see if slide exists in database
    if(new_slide_id %in% dbGetQuery(con, "SELECT slide_id FROM slides")$slide_id){
      print("The slide you are trying to add already exists in the database. Please try
          again and choose 'existing slide'.");
    }else{
      #Split out hole ID from slide ID
      new_hole_id <- strsplit(new_slide_id,",")[[1]][1];
      print("VG (very good) = majority of specimens observed are complete, with spines
          intact, no overgrowths, or recrystallization. Nearly all specimens are
          determinable. \nG (good) = many specimens are complete with spines intact,
          little or no overgrowths, cement or matrix infill present, but outer surface
          intact. Most specimens are determinable. \nM (moderate) = a substantial portion
          of the specimens is broken, with some degree of overgrowth, etching, or
          replacement by minerals other than quartz or pyrite. Fifty percent of specimens
          are determinable. \nP (poor) = specimens are mostly broken and fragmented or
          strongly etched or replaced by other minerals. Less than 5% of specimens are
          determinable. \nVP (very poor) = specimens are only present as inner molds or
          ghosts, or fragments. None are determinable. Estimates of abundance are
          qualitative and were determined on the percentage of radiolarian specimens
          observed in the residue.\n\nPreservation?");
      new_slide_pres <- readline("Preservation?");
      new_slide_depth <- readline("Depth (mbsf)?");
      new_slide_age <- readline("Age (Ma)?");
      #Make insert statement for slide
      sql <- paste("INSERT INTO slides VALUES ('",new_slide_id,"','",new_hole_id,"',",new_
          slide_depth,",",new_slide_age,",'",new_slide_pres,"')",sep="");
      #Send SQLite command
      dbGetQuery(con, sql);
      #Is this a new hole?
      if(!(new_hole_id %in% dbGetQuery(con, "SELECT hole_id FROM holes")$hole_id))
      {
```

Code G.2: RadDataInterface.R (continued)

```
303         #Get lat and long of hole (ask user)
304         new_lat <- readline("Latitude?");
305         new_long <- readline("Longitude?");
306         #Make insert statement for hole
307         sql <- paste("INSERT INTO holes VALUES (",new_hole_id,",",new_lat,",",new_long,")",
                sep="");
308         #Send SQLite command
309         dbGetQuery(con, sql);
310       }
311       #Polite feedback.
312       cat("\nThank you. Slide",new_slide_id,"has been added to RadData.");
313     }
314   }
315   return(new_slide_id);
316 }
317
318
319 ####################################################################
320 #Function to print out the last set of measurements added to the
321 #measurements table of RadData (measurements sharing the highest
322 #indiv_id entry in the table)
323 reviewMeas <- function(con)
324 {
325   ans <- dbGetQuery(con, "SELECT * FROM measurements WHERE indiv_id=(SELECT MAX(indiv_id)
           FROM measurements)");
326   print(ans);
327
328 }
329
330
331
332 ####################################################################
333 #Function to print number of specimens (individuals) in database
334 #total, and number from current slide in database
335 reportTotals <- function(con,slide_id)
336 {
337   cat("Total individuals measured:   ");
338   ans <- max(dbGetQuery(con, "SELECT indiv_id FROM individuals"));
339   cat(ans);
340   cat(paste("\nIndividuals measured on slide "),slide_id,":   ",sep="");
341   query <- paste("SELECT indiv_id FROM individuals WHERE slide_id='",slide_id,"'",sep="");
342   ans <- dbGetQuery(con,query);
343   cat(length(ans$indiv_id));
344   cat(paste("\nIndividuals on that slide by species:   ",sep=""));
345   query <- paste("SELECT species FROM individuals WHERE slide_id='",slide_id,"'",sep="");
346   cat(paste("\n"));
347   ans <- dbGetQuery(con,query);
348   print(table(ans));
349 }
350
351 ####################################################################
352 #Function to create an entry in the individuals table of RadData,
353 #prompt the user to make measurements for that individual in
354 #ImageJ (diplaying the required measurements and order they need
355 #to be in), and collect measurements written by ImageJ to the
356 #pipe.txt file, creating appropriate entries in the measurements
357 #table of RadData.
358 newIndividual <- function(current_sp,slide_id,con)
```

Code G.2: RadDataInterface.R (continued)

```
{
  #Define shorthands for the species name strings
  #Didymocyrtis-Diartus
  La <- 'Lithocyclia angusta';
  Dpr <- 'Didymocyrtis prismatica';
  Dv <- 'Didymocyrtis violina';
  Dl <- 'Didymocyrtis laticonus';
  Dm <- 'Didymocyrtis mammifera';
  Dan <- 'Didymocyrtis antepenultima';
  Dpen <- 'Didymocyrtis penultima';
  Dt <- 'Didymocyrtis tetrathalamus';
  Dav <- 'Didymocyrtis avita';
  Dpet <- 'Diartus petterssoni';
  Dh <- 'Diartus hughesi';
  #Artophormis
  Ab <- 'Artophormis barbadensis';
  Ag <- 'Artophormis gracilis';
  #Stichocorys
  Sd <- 'Stichocorys delmontensis';
  Sp <- 'Stichocorys peregrina';
  Sw <- 'Stichocorys wolffii';
  #Centrobotrys
  Cg <- 'Centrobotrys gravida';
  Cp <- 'Centrobotrys petrushevskayae';
  Ct <- 'Centrobotrys thermophila';
  #Phormocyrtis
  Pse <- 'Phormocyrtis striata exquisita';
  Pss <- 'Phormocyrtis striata striata';
  #Theocyrtis
  Tt <- 'Theocyrtis tuberosa';
  Ta <- 'Theocyrtis annosa';

  cat("To which species does the new individual belong?");
  switch(menu(c(La,Dpr,Dv,Dl,Dm,Dan,Dpen,Dt,Dav,Dpet,Dh,Ab,Ag,Sd,Sp,Sw,Cg,Cp,Ct,Pse,Pss,Tt,
       Ta)) + 1, cat("Nothing done\n"), new_indiv_sp <- La, new_indiv_sp <- Dpr, new_indiv_
       sp <- Dv, new_indiv_sp <- Dl, new_indiv_sp <- Dm, new_indiv_sp <- Dan, new_indiv_sp
       <- Dpen, new_indiv_sp <- Dt, new_indiv_sp <- Dav, new_indiv_sp <- Dpet, new_indiv_sp
        <- Dh, new_indiv_sp <- Ab, new_indiv_sp <- Ag, new_indiv_sp <- Sd, new_indiv_sp <-
       Sp, new_indiv_sp <- Sw, new_indiv_sp <- Cg, new_indiv_sp <- Cp, new_indiv_sp <- Ct,
       new_indiv_sp <- Pse, new_indiv_sp <- Pss, new_indiv_sp <- Tt, new_indiv_sp <- Ta);
  #If the species has changed, display the new measurement pic
  if (new_indiv_sp != current_sp)
  {
    #Pull up the matching measurement pic for the new species
    plot(read.pnm(paste("Images/", new_indiv_sp, ".pgm", sep="")));
  }
  #Ask the user for a comment about the individual
  indiv_comment <- readline("\nComment on this individual: ");
  #Now instruct user to go and make measurements
  dummy <- readline("\nGo make your measurements. When you're done, hit return to continue.
       ");
  #Alright. Here comes the juicy part. Read the measurements
  #from file and turn them into a SQL INSERT statement.
  #First, we need the order of measurements we're expecting
  #given the species at hand.
  measOrder <- getMeasOrder(new_indiv_sp);
  #Now retrieve the measurements from the file, add the IDs for
  #the measurement types from above, and cull out the zero
```

Code G.2: RadDataInterface.R (continued)

```
  #valued measurements
  fileMeas <- readMeasurements(measOrder);
  #Now check to see if the right number of measurements were
  #actually in the –fileif not, tell the user something's wrong
  if(fileMeas[[1]]==FALSE){
    dummy <- readline(paste("\nThere was an error in reading the file. Expected", length(
        measOrder), "measurements in the file! Please fix the file and try again."));
  }else{  #File is OK and ready to be processed
    #Create new entry in the individuals table of RadData
    sql <- paste("INSERT INTO individuals VALUES (NULL,'",slide_id,"'",",'",new_indiv_sp,"
        ','",indiv_comment,"')", sep="");
    dbGetQuery(con,sql);
    #Retrieve primary key column entry for the row just
    #created (i.e. the indiv_ID), for use in creating
    #measurement entries
    current_indiv_ID <- as.numeric(dbGetQuery(con,"SELECT last_insert_rowid()"));
    #Generate a SQL INSERT statement for each of these
    #measurements, one by one
    for(i in 1:nrow(fileMeas[[2]]))
    {
      #Send the ith row out for parsing to SQL
      sql <- createSQLInsert(fileMeas[[2]][i,],current_indiv_ID);
      #Send SQL INSERT statement to RadData.db
      dbGetQuery(con,sql);
      #Print last inserted row to make sure it showed up
      #dbGetQuery(con,"SELECT * FROM measurements WHERE rowid=last_insert_rowid()");
    }
    #Finally, empty the file so it's ready for the next pipe
    write.table(NA,append=FALSE,quote=FALSE,na="",row.names=FALSE,col.names=FALSE,file="
        pipefilename.txt",eol="");
  }

  #In a few cases, additional measurements (counts) are needed that cannot
  #be added in ImageJ, but must be entered manually:
  #(for species Diartus hughesi and most of the Didymocyrtises)
  #For Diartus hughesi, get number of polar caps
  if(new_indiv_sp == Dh)
  {
    V1 <- readline("Number of polar caps?  ");
    V2 <- "No file";
    measOrder <- 55;
    meas <- as.data.frame(cbind(V1,V2,measOrder))
    sql <- createSQLInsert(meas,current_indiv_ID);
  }
  #For the relevant Didymocyrtises, get the number of plicae
  if(new_indiv_sp %in% c(Dl,Dan,Dv,Dm,Dpen,Dav))
  {
    V1 <- readline("Number of tubercles/plicae?  ");
    V2 <- "No file";
    measOrder <- 54;
    meas <- as.data.frame(cbind(V1,V2,measOrder))
    sql <- createSQLInsert(meas,current_indiv_ID);
  }


  #Send back the new species name
  return(new_indiv_sp);
}
```

Code G.2: RadDataInterface.R (continued)

```
465
466
467
468 ######################################################################
469 #Script:
470 ######################################################################
471
472
473
474 #Friendly welcome to the interface
475 cat("\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\nWelcome to the RadData
      database.");
476
477 #Connect to the RadData.sqlite database
478 #Load required libraries
479 library(RSQLite);
480 library(pixmap);
481 #Change working directory to point to the directory where RadData
482 #database file is stored.
483 #setwd('/Users/Ben/Dropbox/Harvard/By-Lineage\ Rads/RadData\ Database');
484 #Open database connection
485 con <- dbConnect(SQLite(), "RadData.sqlite");
486 #Feedback to user
487 cat("\nConnection to RadData.sqlite opened.");
488 #Set current species variable
489 current_sp <- "No species set";
490 #Open a new plot window of the right size
491 quartz(title='RadData',width=5, height=3.5, dpi=100);
492 #Set margins in window to none (pic goes to edge)
493 par(mar=c(0,0,0,0));
494 #Pull up the splashpage
495 plot(read.pnm("Images/Splashpage.pgm"));
496
497 #Set menuActive variable; while TRUE, loop through menu,
498 #when user quits, set to FALSE and stop looping through menu
499 menuActive <- TRUE;
500 while (menuActive == TRUE)
501 {
502   #Present a series of options
503   cat(rep("\n",17),"What would you like to do?\n");
504   switch(menu(c("Set slide", "New individual", "Review last set of measurements added", "
       How many specimens measured so far?", "Save database backup","Exit RadData interface
       ")) + 1, cat("Nothing done\n"), slide_id <- setSlide(con), current_sp <-
       newIndividual(current_sp,slide_id,con), reviewMeas(con), reportTotals(con,slide_id),
        saveBackup(con), menuActive <- goodbye(con));
505 }
```

G.3 R script for analyzing radiolarian data from RadData

Code G.3: AnalyzeRadData.R

```
1 #Code to interrogate the RadData database
2 #Benjamin Kotrc, February 2013
3 #kotrc@fas.harvard.edu
```

Code G.3: AnalyzeRadData.R (continued)

```
4
5  #Load required libraries
6  library(RSQLite)
7  library(reshape)
8  library(maps)
9  library(paleoTS)
10 library(xtable)
11 source("berg95.R")
12
13 #Create a new connection with a database (if the file doesn't
14 #exist yet, it'll be created). Returns handle for database
15 #connection.
16 con <- dbConnect(SQLite(), "RadData.sqlite")
17
18 ############################
19 ###FINAL PLOTS FOR THESIS###
20 ############################
21
22 ###STABILIZING AVERAGES#####
23
24 #Slide of interest:
25 sl_int <- "573,7,6,41-48"
26 #SQL query to retrieve S. peregrina and S. delmontensis
27 meas <- dbGetQuery(con, "SELECT slides.slide_id, individuals.species, individuals.indiv_id,
       measurements.meas_type, measurements.meas_value FROM slides INNER JOIN individuals ON
       slides.slide_id=individuals.slide_id INNER JOIN measurements ON measurements.indiv_id
      =individuals.indiv_id WHERE slides.slide_id='573,7,6,41-48' AND individuals.species IN
       ('Stichocorys peregrina', 'Stichocorys delmontensis') ORDER BY individuals.indiv_id")
28 #Transform data so each row is a specimen, with columns
29 #as measurements
30 #Get rid of species names and slide_id
31 meas <- meas[,3:5]
32 #Equivalent of "melt"
33 names(meas) <- c("id","variable","value");
34 #Now "cast" with rows=individuals and columns=measurement types
35 meas <- cast(meas, id ~ variable)
36 #For reference, here's what the measurement IDs mean:
37 data.frame(Measurement=measName(as.numeric(colnames(meas)[2:12])),row.names=colnames(meas)
      [2:12])
38 #First let's see what the "average" specimen looks like
39 #as the number of measurements piles up
40 #Fill in measurement averages for missing values
41 meas2 <- meas
42 for(i in 2:dim(meas2)[2])
43 {
44   meas2[is.na(meas2[,i]),i] <- mean(meas2[,i],na.rm=TRUE)
45 }
46 #Source external file with geometric functions
47 source("RadShapeFunctions.R")
48 #Calculate silicifications
49 sivals <- getStichoSi(meas2[,2:12])
50 #Plot
51 #hist(sivals)
52 #Get the fonts ready
53 file.exists <- function( fname ) length(Sys.glob(fname))>0
54 absolute.path.to.font.files <-  "/Users/bkotrc/font/";
55 bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
56 ## if you do not have the correct font types
```

Code G.3: AnalyzeRadData.R (continued)

```
for (i in 1:length(bera.names)) {
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".afm", sep="")) )
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".otf", sep="")) )
}
gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files,
                                        bera.names, ".afm", sep=""))
pdfname <- "stabilizingaverages.pdf";
#Make a composite plot of previous plus this plot for publication
pdf(file=pdfname, bg="white", width=(7.2/cm(1)), height=(7.2/cm(1)), pointsize=7, family=
    gillsans);
#Now do the random-order plotting thing
x <- sivals
plot(seq_along(x),cumsum(x)/seq_along(x),pch=16,col="#00000050",ylim=c(0,15),axes=FALSE,
    xlab='Number of Specimens',ylab='Mean Silicification %age',bty="n",main="",xaxs="i",
    xpd=TRUE)
axthck <- 0.3
axis(1, lwd.ticks=axthck, lwd=axthck, at=seq(0,100,10),xpd=TRUE)
axis(2, lwd=axthck)
#Now do this again for a whole bunch of random times
for(i in 1:500)
{
  y <- x[sample(length(x))]
  points(seq_along(y),cumsum(y)/seq_along(y),pch=16,col="#00000010",xpd=TRUE)
}
dev.off();
#Now embed font in that file
embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font");

###END STAB. AVER.#####


###TREND IN STICHOCORYS#####

#SQL query to retrieve S. peregrina and S. delmontensis and
#S. wolffii from all slides
meas <- dbGetQuery(con, "SELECT holes.hole_id, holes.latitude, holes.longitude, slides.age,
     individuals.species, individuals.indiv_id, measurements.meas_type, measurements.meas_
    value FROM holes INNER JOIN slides ON holes.hole_id=slides.hole_id  INNER JOIN
    individuals ON slides.slide_id=individuals.slide_id INNER JOIN measurements ON
    measurements.indiv_id=individuals.indiv_id WHERE individuals.species IN ('Stichocorys
    peregrina', 'Stichocorys delmontensis','Stichocorys wolffii') ORDER BY individuals.
    indiv_id")

#Now, let's melt-cast this so that each row is a specimen so we can get silica %ages
names(meas) <- c("hole_id","latitude","longitude","age","species","indiv_id","variable","
    value")
meas <- cast(data=meas)

#In order to calculate silicifications, we need the geometric functions
source("RadShapeFunctions.R")
#Now calculate silicifications
meas$sipct <- getStichoSi(meas)
#Need to fill in "average"
#values for missing data.
#These are the slide ages we have
slides <- unique(meas$age)
```

Code G.3: AnalyzeRadData.R (continued)

```
#Set up a variable to hold the average silicification for each slide
meansipct <- slides
#And another one to keep track of which hole each slide is from
slidehole <- slides
#Loop through each slide
for(i in 1:length(slides))
{
  #Extract the individuals from the current slide
  temp <- meas[as.numeric(meas$age) == slides[i],]
  #Replace NAs with average values for that slide
  for(j in 7:17)
  {
    temp[is.na(temp[,j]),j] <- mean(temp[,j],na.rm=TRUE)
  }
  #now we need to calculate the silicification percentages again
  temp$sipct <- getStichoSi(temp)
  #Now copy that extract back into meas
  meas[as.numeric(meas$age) == slides[i],] <- temp
  #While we're at it, let's calculate the average silicification for each slide
  meansipct[i] <- mean(temp$sipct)
  #And keep track of the hole this slide is from
  slidehole[i] <- unique(temp$hole_id)
}
#Reorder the mean values so that they plot in the right order
x <- slides
y <- meansipct
z <- slidehole
means <- data.frame(x,y,z)
means <- means[order(means$x),]

#Now plot
#Get the fonts ready
file.exists <- function( fname ) length(Sys.glob(fname))>0
absolute.path.to.font.files <-  "/Users/bkotrc/font/";
bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
## if you do not have the correct font types
for (i in 1:length(bera.names)) {
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".afm", sep="")) )
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".otf", sep="")) )
}
gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files,
                                        bera.names, ".afm", sep=""))
pdfname <- "stichotrend2.pdf"
pdf(file=pdfname, bg="white", width=(7.2/cm(1)), height=(7.2/cm(1)), pointsize=7, family=
    gillsans);

#No points at first, we want them over the geologic timescale
axthck <- 0.3;
par(mar=c(4.1,4.1,1.1,1.1))
plot(meas$age,meas$sipct,xlim=c(25,0),ylim=c(0,16),xlab="Geologic Time (Ma)", ylab="%
    Silicification",pch=16,col="#70707050",type="n",bty="n",main="",xaxs="i",yaxs="i",xpd=
    TRUE,axes=FALSE)
#Add axes manually
axis(1, lwd=axthck)
axis(2, lwd=axthck)
berg95(start='mio',line=axthck)
```

Code G.3: AnalyzeRadData.R (continued)

```
#Now the points
points(meas$age,meas$sipct,pch=c(16,16,17)[as.numeric(as.factor(meas$hole_id))],col="
     #70707050",xpd=TRUE)
#Add average values
points(means$x,means$y,cex=1.25,pch=c(16,16,17)[as.numeric(as.factor(means$z))],type="b")
#Close connection to pdf file
dev.off()
#Now embed font in that file
embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font")

#Separate file for associated map of drill sites
#Do this —manuallyPDF export didn't work for that
#Composite this with the other figure manually
par(mar=c(0.1,0.1,0.1,0.1))
map(database="world2", interior=FALSE, col="black", xlim=c(100,300),ylim=c(-60,60))
box()
points(360+unique(meas$longitude),unique(meas$latitude),cex=2,pch=c(16,17))
text(360+unique(meas$longitude),unique(meas$latitude),labels=c("DSDP 573","DSDP 71"),pos=4,
     cex=.8)

###END TR. IN STICHOCORYS#####


###TREND IN DIDYMOCYRTIS#####

#Cool beans. Now, retry this for the Didymocyrtis lineage
#Again, get the data from RadData.sqlite first
meas <- dbGetQuery(con, "SELECT holes.hole_id, holes.latitude, holes.longitude, slides.age,
      individuals.species, individuals.indiv_id, measurements.meas_type, measurements.meas_
     value FROM holes INNER JOIN slides ON holes.hole_id=slides.hole_id  INNER JOIN
     individuals ON slides.slide_id=individuals.slide_id INNER JOIN measurements ON
     measurements.indiv_id=individuals.indiv_id WHERE individuals.species LIKE '
     Didymocyrtis%' ORDER BY individuals.indiv_id")

#Now, let's melt-cast this so that each row is a specimen so we can get silica %ages
names(meas) <- c("hole_id","latitude","longitude","age","species","indiv_id","variable","
     value")
meas <- cast(data=meas)

#These are the slide ages we have
slides <- unique(meas$age)
#Set up a variable to hold the average silicification for each slide
meansipct <- slides
#And another one to keep track of which hole each slide is from
slidehole <- slides

#In order to calculate silicifications, we need the geometric functions
source("RadShapeFunctions.R")

#The Didymocyrtis species, grouped by geometric model
spgroups <- list("Didymocyrtis tetrathalamus","Didymocyrtis prismatica",c("Didymocyrtis
     penultima","Didymocyrtis antepenultima", "Didymocyrtis avita", "Didymocyrtis laticonus
     "),c("Didymocyrtis violina","Didymocyrtis mammifera"))

meas$sipct <- NA

#Loop through each slide
for(i in 1:length(slides))
```

Code G.3: AnalyzeRadData.R (continued)

```
{
  #Also, loop through each set of species (grouped by geometric model)
  for(a in 1:length(spgroups))
  {
    #Extract the individuals from the current slide and current species set
    temp <- meas[(as.numeric(meas$age) == slides[i]) & (meas$species %in% spgroups[[a]]),]
    #If the extract isn't empty (i.e. if there are matching specimens)
    if(dim(temp)[1] > 0)
    {
      #Replace NAs with average values for that slide and current species set
      for(j in 7:26)
      {
        #Only do it if there are non-NA values in that column
        if(sum(is.na(temp[,j])) > 0)
        {
          temp[is.na(temp[,j]),j] <- mean(temp[,j],na.rm=TRUE)
        }
      }
      #Now calculate the silicification percentages
      if(a == 1){temp$sipct <- getDtetraSi(temp)}
      if(a == 2){temp$sipct <- getDprisSi(temp)}
      if(a == 3){temp$sipct <- getDpenSi(temp)}
      if(a == 4){temp$sipct <- getDvioSi(temp)}
      #Now copy that extract back into meas
      meas[(as.numeric(meas$age) == slides[i]) & (meas$species %in% spgroups[[a]]),] <-
           temp
    }
  }
  #Now extract all specimens from this slide
  temp <- meas[(as.numeric(meas$age) == slides[i]),]
  #Calculate the average silicification for this slide
  meansipct[i] <- mean(temp$sipct)
  #And keep track of the hole this slide is from
  slidehole[i] <- unique(temp$hole_id)
}

#Reorder the mean values so that they plot in the right order
x <- slides
y <- meansipct
z <- slidehole
means <- data.frame(x,y,z)
means <- means[order(means$x),]

#Now plot
#Get the fonts ready
file.exists <- function( fname ) length(Sys.glob(fname))>0
absolute.path.to.font.files <-  "/Users/bkotrc/font/";
bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
## if you do not have the correct font types
for (i in 1:length(bera.names)) {
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".afm", sep="")) )
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".otf", sep="")) )
}
gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files,
                                        bera.names, ".afm", sep=""))
pdfname <- "didymotrend.pdf"
```

Code G.3: AnalyzeRadData.R (continued)

```
pdf(file=pdfname, bg="white", width=(7.2/cm(1)), height=(7.2/cm(1)), pointsize=7, family=
    gillsans);

#No points at first, we want them over the geologic timescale
axthck <- 0.3;
par(mar=c(4.1,4.1,1.1,1.1))
plot(meas$age,meas$sipct,xlim=c(25,0),ylim=c(0,25),xlab="Geologic Time (Ma)", ylab="%
    Silicification",pch=16,col="#70707050",type="n",bty="n",main="",xaxs="i",yaxs="i",xpd=
    TRUE,axes=FALSE)
#Add axes manually
axis(1, lwd=axthck)
axis(2, lwd=axthck)
berg95(start='mio',line=axthck)
#Now the points
points(meas$age,meas$sipct,pch=c(16,16,17)[as.numeric(as.factor(meas$hole_id))],col="
    #70707050",xpd=TRUE)
#Add average values
points(means$x,means$y,cex=1.25,pch=c(16,16,17)[as.numeric(as.factor(means$z))],type="b")
#Close connection to pdf file
dev.off()
#Now embed font in that file
embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font")

###END TR. IN DIDYMOCYRTIS#####


###TREND IN CENTROBOTRYS#####

#Again, get the data from RadData.sqlite first
meas <- dbGetQuery(con, "SELECT holes.hole_id, holes.latitude, holes.longitude, slides.age,
     individuals.species, individuals.indiv_id, measurements.meas_type, measurements.meas_
    value FROM holes INNER JOIN slides ON holes.hole_id=slides.hole_id  INNER JOIN
    individuals ON slides.slide_id=individuals.slide_id INNER JOIN measurements ON
    measurements.indiv_id=individuals.indiv_id WHERE individuals.species LIKE '
    Centrobotrys%' ORDER BY individuals.indiv_id")

#Now, let's melt-cast this so that each row is a specimen so we can get silica %ages
names(meas) <- c("hole_id","latitude","longitude","age","species","indiv_id","variable","
    value")
meas <- cast(data=meas)

#These are the slide ages we have
slides <- unique(meas$age)
#Set up a variable to hold the average silicification for each slide
meansipct <- slides
#And another one to keep track of which hole each slide is from
slidehole <- slides

#In order to calculate silicifications, we need the geometric functions
source("RadShapeFunctions.R")

#The Didymocyrtis species, grouped by geometric model
spgroups <- list("Centrobotrys gravida","Centrobotrys petrushevskayae","Centrobotrys
    thermophila")

meas$sipct <- NA

#Loop through each slide
```

Code G.3: AnalyzeRadData.R (continued)

```
for(i in 1:length(slides))
{
  #Also, loop through each set of species (grouped by geometric model)
  for(a in 1:length(spgroups))
  {
    #Extract the individuals from the current slide and current species set
    temp <- meas[(as.numeric(meas$age) == slides[i]) & (meas$species %in% spgroups[[a]]),]
    #If the extract isn't empty (i.e. if there are matching specimens)
    if(dim(temp)[1] > 0)
    {
      #Replace NAs with average values for that slide and current species set
      for(j in 7:19)
      {
        #Only do it if there are non-NA values in that column
        if(sum(is.na(temp[,j])) < length(temp[,j]))
        {
          temp[is.na(temp[,j]),j] <- mean(temp[,j],na.rm=TRUE)
        }
      }
      cat(a)
      #Now calculate the silicification percentages
      if(a == 1){temp$sipct <- getCgravSi(temp)}
      if(a == 2){temp$sipct <- getCpetSi(temp)}
      if(a == 3){temp$sipct <- getCthermSi(temp)}
      #Now copy that extract back into meas
      meas[(as.numeric(meas$age) == slides[i]) & (meas$species %in% spgroups[[a]]),] <-
           temp
    }
  }
  #Now extract all specimens from this slide
  temp <- meas[(as.numeric(meas$age) == slides[i]),]
  #Calculate the average silicification for this slide
  meansipct[i] <- mean(temp$sipct)
  #And keep track of the hole this slide is from
  slidehole[i] <- unique(temp$hole_id)
}

#Reorder the mean values so that they plot in the right order
x <- slides
y <- meansipct
z <- slidehole
means <- data.frame(x,y,z)
means <- means[order(means$x),]

#Now plot
#Get the fonts ready
file.exists <- function( fname ) length(Sys.glob(fname))>0
absolute.path.to.font.files <-  "/Users/bkotrc/font/";
bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
## if you do not have the correct font types
for (i in 1:length(bera.names)) {
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".afm", sep="")) )
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".otf", sep="")) )
}
gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files,
                                        bera.names, ".afm", sep=""))
```

Code G.3: AnalyzeRadData.R (continued)

```
pdfname <- ”centrotrend.pdf”
pdf(file=pdfname, bg=”white”, width=(7.2/cm(1)), height=(7.2/cm(1)), pointsize=7, family=
    gillsans);

#No points, we want them over the geologic timescale
axthck <- 0.3;
par(mar=c(4.1,4.1,1.1,1.1))
plot(meas$age,meas$sipct,xlim=c(35,0),ylim=c(0,25),xlab=”Geologic Time (Ma)”, ylab=”%
    Silicification”,pch=16,col=”#70707050”,type=”n”,bty=”n”,main=””,xaxs=”i”,yaxs=”i”,xpd=
    TRUE,axes=FALSE)
#Add axes manually
axis(1, lwd=axthck)
axis(2, lwd=axthck)
berg95(line=axthck)
#Now the points
points(meas$age,meas$sipct,pch=c(18,15,16,16,17)[as.numeric(as.factor(meas$hole_id))],col=”
    #70707050”,xpd=TRUE)
#Add average values
#Symbols as follows (hole - factor level - pch - symbol):
#573 - 3 - 16 - circle
#71 - 5 - 17 - triangle
#573B - 4 - 16 - circle
#289 - 2 - 15 - square
#162 - 1 - 18 - diamond
points(means$x,means$y,cex=1.25,pch=c(18,15,16,16,17)[as.numeric(as.factor(means$z))],type=
    ”b”)

#Close connection to pdf file
dev.off()
#Now embed font in that file
embedFonts(file=pdfname,outfile=pdfname, fontpaths=”/Users/bkotrc/font”)

#Separate file for associated map of drill sites
#Do this –manuallyPDF export didn’t work for that
#Composite this with the other figure manually
meas <- dbGetQuery(con,”SELECT * FROM holes”)[1:5,]
#Fix the one value that was entered incorrectly
meas[3,’latitude’] <- 14.8698333
par(mar=c(0.1,0.1,0.1,0.1))
map(database=”world2”, interior=FALSE, col=”black”, xlim=c(100,300),ylim=c(-60,60))
box()
points((meas$longitude %% 360),meas$latitude,cex=2,pch=c(16,16,18,15,17))
text((meas$longitude %% 360),meas$latitude,labels=c(”DSDP 573”,”DSDP 573”,”DSDP 162”,”DSDP
    289”,”DSDP 71”),pos=4,cex=.8)

###END TR. IN CENTROBOTRYS#####


###THICKNESS VS. POROSITY######

#Awesome. Now try and see if we can relate thickness and porosity at all
meas <- dbGetQuery(con, ”SELECT holes.hole_id, holes.latitude, holes.longitude, slides.age,
     individuals.species, individuals.indiv_id, measurements.meas_type, measurements.meas_
    value FROM holes INNER JOIN slides ON holes.hole_id=slides.hole_id  INNER JOIN
    individuals ON slides.slide_id=individuals.slide_id INNER JOIN measurements ON
    measurements.indiv_id=individuals.indiv_id ORDER BY individuals.indiv_id”)

#Now, let’s melt-cast this so that each row is a specimen so we can get silica %ages
```

Code G.3: AnalyzeRadData.R (continued)

```
names(meas) <- c("hole_id","latitude","longitude","age","species","indiv_id","variable","value")
meas <- cast(data=meas)

#All the measurement identifiers explained!
data.frame(names(meas)[7:50],measName(as.numeric(names(meas)[7:50])))

#Plot individually
#plot(meas$'38',meas$'46',xlim=c(0,1),xlab="Pore area",ylab="Shell thickness")
#points(meas$'43',meas$'49',pch=16)
#points(meas$'44',meas$'50',pch=16,col="#00000050")
#points(meas$'45',meas$'51',pch=17,col="#50000050")

#Plot together
poresnwalls <- data.frame(meas$'38',meas$'46')
names(poresnwalls) <- c("Pore area","Shell thickness")
two <- data.frame(meas$'43',meas$'49')
names(two) <- c("Pore area","Shell thickness")
three <- data.frame(meas$'44',meas$'50')
names(three) <- c("Pore area","Shell thickness")
four <- data.frame(meas$'45',meas$'51')
names(four) <- c("Pore area","Shell thickness")
poresnwalls <- rbind(poresnwalls,two,three,four)
#Ignore spurious data
poresnwalls <- poresnwalls[poresnwalls$"Pore area" <= 1,]

#Now plot
#Get the fonts ready
file.exists <- function( fname ) length(Sys.glob(fname))>0
absolute.path.to.font.files <-  "/Users/bkotrc/font/";
bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
## if you do not have the correct font types
for (i in 1:length(bera.names)) {
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".afm", sep="")) )
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".otf", sep="")) )
}
gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files,
                                        bera.names, ".afm", sep=""))
pdfname <- "poresnwalls.pdf"
pdf(file=pdfname, bg="white", width=(7.2/cm(1)), height=(7.2/cm(1)), pointsize=7, family=gillsans);

poresnwalls[,1] <- poresnwalls[,1]*100

#No points, we want them over the geologic timescale
axthck <- 0.3;
par(mar=c(4.1,4.1,1.1,1.1))
plot(poresnwalls,xlab="Pore area (%)",ylab="Shell thickness (µm)",xlim=c(0,100),ylim=c(0,10),pch=16,col="#70707050",bty="n",main="",xaxs="i",yaxs="i",xpd=TRUE,axes=FALSE)
#Add axes manually
axis(1, lwd=axthck)
axis(2, lwd=axthck)

#Close connection to pdf file
dev.off()
#Now embed font in that file
```

Code G.3: AnalyzeRadData.R (continued)

```
embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font")

###END THICKNESS/POROSITY##


###STICHOCORYS HISTOGRAMS###

#SQL query to retrieve S. peregrina and S. delmontensis and
#S. wolffii from all slides
meas <- dbGetQuery(con, "SELECT holes.hole_id, holes.latitude, holes.longitude, slides.age,
     individuals.species, individuals.indiv_id, measurements.meas_type, measurements.meas_
     value FROM holes INNER JOIN slides ON holes.hole_id=slides.hole_id  INNER JOIN
     individuals ON slides.slide_id=individuals.slide_id INNER JOIN measurements ON
     measurements.indiv_id=individuals.indiv_id WHERE individuals.species IN ('Stichocorys
     peregrina', 'Stichocorys delmontensis','Stichocorys wolffii') ORDER BY individuals.
     indiv_id")

#Now, let's melt-cast this so that each row is a specimen so we can get silica %ages
names(meas) <- c("hole_id","latitude","longitude","age","species","indiv_id","variable","
     value")
meas <- cast(data=meas)

#In order to calculate silicifications, we need the geometric functions
source("RadShapeFunctions.R")
#Now calculate silicifications
meas$sipct <- getStichoSi(meas)
#OK, works fine--but does not handle NAs in the data. Need to fill in "average"
#values for missing data.
#Get the measurement names we're talking about
#source("RadDataInterface.R")
#measName(as.numeric(names(meas)[7:17]))
#These are the slide ages we have
slides <- unique(meas$age)
#Set up a variable to hold the average silicification for each slide
meansipct <- slides
#Set up a variable to hold the variance in silicification for each slide
varsipct <- slides
#And another one to keep track of sample size in each slide
nsamp <- slides
#Loop through each slide
for(i in 1:length(slides))
{
  #Extract the individuals from the current slide
  temp <- meas[as.numeric(meas$age) == slides[i],]
  #Replace NAs with average values for that slide
  for(j in 7:17)
  {
    temp[is.na(temp[,j]),j] <- mean(temp[,j],na.rm=TRUE)
  }
  #now we need to calculate the silicification percentages again
  temp$sipct <- getStichoSi(temp)
  #Now copy that extract back into meas
  meas[as.numeric(meas$age) == slides[i],] <- temp
  #While we're at it, let's calculate the average silicification for each slide
  meansipct[i] <- mean(temp$sipct)
  #And also the variance
  varsipct[i] <- var(temp$sipct)
  #Keep track of sample size
```

Code G.3: AnalyzeRadData.R (continued)

```
  nsamp[i] <- dim(temp)[1]
  #And keep track of the hole this slide is from
  slidehole[i] <- unique(temp$hole_id)
}
#Get the fonts ready
file.exists <- function( fname ) length(Sys.glob(fname))>0
absolute.path.to.font.files <-  "/Users/bkotrc/font/";
bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
## if you do not have the correct font types
for (i in 1:length(bera.names)) {
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".afm", sep="")) )
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".otf", sep="")) )
}
gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files,
                                        bera.names, ".afm", sep=""))
pdfname <- "stichohists.pdf"
pdf(file=pdfname, bg="white", width=(7.2/cm(1)), height=(7.2/cm(1)), pointsize=7, family=
    gillsans);
#Plot histograms
par(mar=c(4.1,4.1,2.1,2.1),mfrow=c(3,1))
hist(meas$sipct[meas$species == 'Stichocorys peregrina'],breaks=(0:14),main="S. peregrina",
    xlab="% Silicification",col="#757575",font.main=1,axes=FALSE)
axthck <- 0.3;
#Add axes manually
axis(1, lwd=axthck)
axis(2, lwd=axthck)

hist(meas$sipct[meas$species == 'Stichocorys wolffii'],breaks=(0:14),main="S. wolffii",xlab
    ="% Silicification",col="#757575",font.main=1,axes=FALSE)
axthck <- 0.3;
#Add axes manually
axis(1, lwd=axthck)
axis(2, lwd=axthck)

hist(meas$sipct[meas$species == 'Stichocorys delmontensis'],breaks=(0:14),main="S.
    delmontensis",xlab="% Silicification",col="#757575",font.main=1,axes=FALSE)
#Add axes manually
axis(1, lwd=axthck)
axis(2, lwd=axthck)

#Close connection to pdf file
dev.off()
#Now embed font in that file
embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font")

###END STICHO. HIST.########


###DIDYMOCYRTIS HISTOGRAMS###

#Again, get the data from RadData.sqlite first
meas <- dbGetQuery(con, "SELECT holes.hole_id, holes.latitude, holes.longitude, slides.age,
     individuals.species, individuals.indiv_id, measurements.meas_type, measurements.meas_
    value FROM holes INNER JOIN slides ON holes.hole_id=slides.hole_id  INNER JOIN
    individuals ON slides.slide_id=individuals.slide_id INNER JOIN measurements ON
    measurements.indiv_id=individuals.indiv_id WHERE individuals.species LIKE '
```

Code G.3: AnalyzeRadData.R (continued)

```
     Didymocyrtis%' ORDER BY individuals.indiv_id")

#Now, let's melt-cast this so that each row is a specimen so we can get silica %ages
names(meas) <- c("hole_id","latitude","longitude","age","species","indiv_id","variable","
     value")
meas <- cast(data=meas)

#These are the slide ages we have
slides <- unique(meas$age)
#Set up a variable to hold the average silicification for each slide
meansipct <- slides
#And another one to keep track of which hole each slide is from
slidehole <- slides

#In order to calculate silicifications, we need the geometric functions
source("RadShapeFunctions.R")

#The Didymocyrtis species, grouped by geometric model
spgroups <- list("Didymocyrtis tetrathalamus","Didymocyrtis prismatica",c("Didymocyrtis
     penultima","Didymocyrtis antepenultima", "Didymocyrtis avita", "Didymocyrtis laticonus
     "),c("Didymocyrtis violina","Didymocyrtis mammifera"))

meas$sipct <- NA

#Loop through each slide
for(i in 1:length(slides))
{
  #Also, loop through each set of species (grouped by geometric model)
  for(a in 1:length(spgroups))
  {
    #Extract the individuals from the current slide and current species set
    temp <- meas[(as.numeric(meas$age) == slides[i]) & (meas$species %in% spgroups[[a]]),]
    #If the extract isn't empty (i.e. if there are matching specimens)
    if(dim(temp)[1] > 0)
    {
      #Replace NAs with average values for that slide and current species set
      for(j in 7:26)
      {
        #Only do it if there are non-NA values in that column
        if(sum(is.na(temp[,j])) > 0)
        {
          temp[is.na(temp[,j]),j] <- mean(temp[,j],na.rm=TRUE)
        }
      }
      #Now calculate the silicification percentages
      if(a == 1){temp$sipct <- getDtetraSi(temp)}
      if(a == 2){temp$sipct <- getDprisSi(temp)}
      if(a == 3){temp$sipct <- getDpenSi(temp)}
      if(a == 4){temp$sipct <- getDvioSi(temp)}
      #Now copy that extract back into meas
      meas[(as.numeric(meas$age) == slides[i]) & (meas$species %in% spgroups[[a]]),] <-
           temp
    }
  }
  #Now extract all specimens from this slide
  temp <- meas[(as.numeric(meas$age) == slides[i]),]
  #Calculate the average silicification for this slide
  meansipct[i] <- mean(temp$sipct)
```

Code G.3: AnalyzeRadData.R (continued)

```
  #And keep track of the hole this slide is from
  slidehole[i] <- unique(temp$hole_id)
}

#Get the fonts ready
file.exists <- function( fname ) length(Sys.glob(fname))>0
absolute.path.to.font.files <-  "/Users/bkotrc/font/";
bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
## if you do not have the correct font types
for (i in 1:length(bera.names)) {
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".afm", sep="")) )
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".otf", sep="")) )
}
gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files,
                                        bera.names, ".afm", sep=""))
pdfname <- "didymohists.pdf"
pdf(file=pdfname, bg="white", width=(7.2/cm(1)), height=(14.4/cm(1)), pointsize=7, family=
    gillsans);

#Species we have; one for each histogram
splist <- c("Didymocyrtis prismatica","Didymocyrtis violina","Didymocyrtis mammifera","
    Didymocyrtis laticonus","Didymocyrtis penultima","Didymocyrtis avita","Didymocyrtis
    tetrathalamus")
splist <- rev(splist)
#Plot histograms
par(mar=c(4.1,4.1,2.1,2.1),mfrow=c(7,1))
#Loop through the species
for (i in 1:length(splist))
{
  hist(meas$sipct[meas$species == splist[i]],breaks=seq(from=0,to=25,by=2.5),main=splist[i
      ],xlab="% Silicification",col="#757575",font.main=1,axes=FALSE)
  axthck <- 0.3;
  #Add axes manually
  axis(1, lwd=axthck)
  axis(2, lwd=axthck)
}
#Close connection to pdf file
dev.off()
#Now embed font in that file
embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font")

###END DIDYMO. HIST.########


###CENTROBOTRYS HISTOGRAMS###

#Again, get the data from RadData.sqlite first
meas <- dbGetQuery(con, "SELECT holes.hole_id, holes.latitude, holes.longitude, slides.age,
     individuals.species, individuals.indiv_id, measurements.meas_type, measurements.meas_
    value FROM holes INNER JOIN slides ON holes.hole_id=slides.hole_id  INNER JOIN
    individuals ON slides.slide_id=individuals.slide_id INNER JOIN measurements ON
    measurements.indiv_id=individuals.indiv_id WHERE individuals.species LIKE '
    Centrobotrys%' ORDER BY individuals.indiv_id")

#Now, let's melt-cast this so that each row is a specimen so we can get silica %ages
```

Code G.3: AnalyzeRadData.R (continued)

```
names(meas) <- c("hole_id","latitude","longitude","age","species","indiv_id","variable","value")
meas <- cast(data=meas)

#These are the slide ages we have
slides <- unique(meas$age)
#Set up a variable to hold the average silicification for each slide
meansipct <- slides
#And another one to keep track of which hole each slide is from
slidehole <- slides

#In order to calculate silicifications, we need the geometric functions
source("RadShapeFunctions.R")

#The Didymocyrtis species, grouped by geometric model
spgroups <- list("Centrobotrys gravida","Centrobotrys petrushevskayae","Centrobotrys thermophila")

meas$sipct <- NA

#Loop through each slide
for(i in 1:length(slides))
{
  #Also, loop through each set of species (grouped by geometric model)
  for(a in 1:length(spgroups))
  {
    #Extract the individuals from the current slide and current species set
    temp <- meas[(as.numeric(meas$age) == slides[i]) & (meas$species %in% spgroups[[a]]),]
    #If the extract isn't empty (i.e. if there are matching specimens)
    if(dim(temp)[1] > 0)
    {
      #Replace NAs with average values for that slide and current species set
      for(j in 7:19)
      {
        #Only do it if there are non-NA values in that column
        if(sum(is.na(temp[,j])) < length(temp[,j]))
        {
          temp[is.na(temp[,j]),j] <- mean(temp[,j],na.rm=TRUE)
        }
      }
      cat(a)
      #Now calculate the silicification percentages
      if(a == 1){temp$sipct <- getCgravSi(temp)}
      if(a == 2){temp$sipct <- getCpetSi(temp)}
      if(a == 3){temp$sipct <- getCthermSi(temp)}
      #Now copy that extract back into meas
      meas[(as.numeric(meas$age) == slides[i]) & (meas$species %in% spgroups[[a]]),] <-
           temp
    }
  }
  #Now extract all specimens from this slide
  temp <- meas[(as.numeric(meas$age) == slides[i]),]
  #Calculate the average silicification for this slide
  meansipct[i] <- mean(temp$sipct)
  #And keep track of the hole this slide is from
  slidehole[i] <- unique(temp$hole_id)
}

```

Code G.3: AnalyzeRadData.R (continued)

```
#Get the fonts ready
file.exists <- function( fname ) length(Sys.glob(fname))>0
absolute.path.to.font.files <-  "/Users/bkotrc/font/";
bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
## if you do not have the correct font types
for (i in 1:length(bera.names)) {
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".afm", sep="")) )
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".otf", sep="")) )
}
gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files,
                                        bera.names, ".afm", sep=""))
pdfname <- "centrohists.pdf"
pdf(file=pdfname, bg="white", width=(7.2/cm(1)), height=(7.2/cm(1)), pointsize=7, family=
    gillsans);

#Species we have; one for each histogram
splist <- c("Centrobotrys gravida","Centrobotrys petrushevskayae","Centrobotrys thermophila
    ")
splist <- rev(splist)
#Plot histograms
par(mar=c(4.1,4.1,2.1,2.1),mfrow=c(3,1))
#Loop through the species
for (i in 1:length(splist))
{
  hist(meas$sipct[meas$species == splist[i]],breaks=seq(from=0,to=25,by=2.5),main=splist[i
      ],xlab="% Silicification",col="#757575",font.main=1,axes=FALSE)
  axthck <- 0.3;
  #Add axes manually
  axis(1, lwd=axthck)
  axis(2, lwd=axthck)

}
#Close connection to pdf file
dev.off()
#Now embed font in that file
embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font")

###END CENTRO. HIST.########


###GRAND COMPARISON PLOT###

#Get data from PNAS paper
pnas <- read.table(file="PNASdata.csv",sep=",",header=TRUE)
#Get data for three lineages
#SQL query to retrieve S. peregrina and S. delmontensis and
#S. wolffii from all slides
meas <- dbGetQuery(con, "SELECT holes.hole_id, holes.latitude, holes.longitude, slides.age,
     individuals.species, individuals.indiv_id, measurements.meas_type, measurements.meas_
    value FROM holes INNER JOIN slides ON holes.hole_id=slides.hole_id  INNER JOIN
    individuals ON slides.slide_id=individuals.slide_id INNER JOIN measurements ON
    measurements.indiv_id=individuals.indiv_id WHERE individuals.species IN ('Stichocorys
    peregrina', 'Stichocorys delmontensis','Stichocorys wolffii') ORDER BY individuals.
    indiv_id")

#Now, let's melt-cast this so that each row is a specimen so we can get silica %ages
```

Code G.3: AnalyzeRadData.R (continued)

```
names(meas) <- c("hole_id","latitude","longitude","age","species","indiv_id","variable","value")
meas <- cast(data=meas)

#In order to calculate silicifications, we need the geometric functions
source("RadShapeFunctions.R")
#Now calculate silicifications
meas$sipct <- getStichoSi(meas)
#OK, works fine--but does not handle NAs in the data. Need to fill in "average"
#values for missing data.
#Get the measurement names we're talking about
#source("RadDataInterface.R")
#measName(as.numeric(names(meas)[7:17]))
#These are the slide ages we have
slides <- unique(meas$age)
#Set up a variable to hold the average silicification for each slide
meansipct <- slides
#Set up a variable to hold the variance in silicification for each slide
varsipct <- slides
#And another one to keep track of sample size in each slide
nsamp <- slides
#Loop through each slide
for(i in 1:length(slides))
{
  #Extract the individuals from the current slide
  temp <- meas[as.numeric(meas$age) == slides[i],]
  #Replace NAs with average values for that slide
  for(j in 7:17)
  {
    temp[is.na(temp[,j]),j] <- mean(temp[,j],na.rm=TRUE)
  }
  #now we need to calculate the silicification percentages again
  temp$sipct <- getStichoSi(temp)
  #Now copy that extract back into meas
  meas[as.numeric(meas$age) == slides[i],] <- temp
  #While we're at it, let's calculate the average silicification for each slide
  meansipct[i] <- mean(temp$sipct)
  #And also the variance
  varsipct[i] <- var(temp$sipct)
  #Keep track of sample size
  nsamp[i] <- dim(temp)[1]
  #And keep track of the hole this slide is from
  #slidehole[i] <- unique(temp$hole_id)
}
#Reorder the mean values so that they plot in the right order
age <- slides
simean <- meansipct
sivar <- varsipct
TSdata <- data.frame(simean,sivar,nsamp,age)
TSdata <- TSdata[order(TSdata$age),]

stichoTS <- as.paleoTS(mm=TSdata$simean,vv=TSdata$sivar,nn=TSdata$nsamp,tt=TSdata$age,oldest="last")
stichosi <- simean
stichoage <- TSdata$age

#Model fitting exercise (following Hunt 2006):
fit3models(stichoTS)
```

Code G.3: AnalyzeRadData.R (continued)

```
834
835 #Again, get the data from RadData.sqlite first
836 meas <- dbGetQuery(con, "SELECT holes.hole_id, holes.latitude, holes.longitude, slides.age,
      individuals.species, individuals.indiv_id, measurements.meas_type, measurements.meas_
     value FROM holes INNER JOIN slides ON holes.hole_id=slides.hole_id  INNER JOIN
     individuals ON slides.slide_id=individuals.slide_id INNER JOIN measurements ON
     measurements.indiv_id=individuals.indiv_id WHERE individuals.species LIKE '
     Didymocyrtis%' ORDER BY individuals.indiv_id")
837
838 #Now, let's melt-cast this so that each row is a specimen so we can get silica %ages
839 names(meas) <- c("hole_id","latitude","longitude","age","species","indiv_id","variable","
     value")
840 meas <- cast(data=meas)
841
842 #These are the slide ages we have
843 slides <- unique(meas$age)
844 #Set up a variable to hold the average silicification for each slide
845 meansipct <- slides
846 #Set up a variable to hold the variance in silicification for each slide
847 varsipct <- slides
848 #And another one to keep track of sample size in each slide
849 nsamp <- slides
850
851 #In order to calculate silicifications, we need the geometric functions
852 source("RadShapeFunctions.R")
853
854 #The Didymocyrtis species, grouped by geometric model
855 spgroups <- list("Didymocyrtis tetrathalamus","Didymocyrtis prismatica",c("Didymocyrtis
     penultima","Didymocyrtis antepenultima", "Didymocyrtis avita", "Didymocyrtis laticonus
     "),c("Didymocyrtis violina","Didymocyrtis mammifera"))
856
857 meas$sipct <- NA
858
859 #Loop through each slide
860 for(i in 1:length(slides))
861 {
862   #Also, loop through each set of species (grouped by geometric model)
863   for(a in 1:length(spgroups))
864   {
865     #Extract the individuals from the current slide and current species set
866     temp <- meas[(as.numeric(meas$age) == slides[i]) & (meas$species %in% spgroups[[a]]),]
867     #If the extract isn't empty (i.e. if there are matching specimens)
868     if(dim(temp)[1] > 0)
869     {
870       #Replace NAs with average values for that slide and current species set
871       for(j in 7:26)
872       {
873         #Only do it if there are non-NA values in that column
874         (sum(is.na(temp[,j])) < length(temp[,j]))
875         {
876           temp[is.na(temp[,j]),j] <- mean(temp[,j],na.rm=TRUE)
877         }
878       }
879       #Now calculate the silicification percentages
880       if(a == 1){temp$sipct <- getDtetraSi(temp)}
881       if(a == 2){temp$sipct <- getDprisSi(temp)}
882       if(a == 3){temp$sipct <- getDpenSi(temp)}
883       if(a == 4){temp$sipct <- getDvioSi(temp)}
```

Code G.3: AnalyzeRadData.R (continued)

```
      #Now copy that extract back into meas
      meas[(as.numeric(meas$age) == slides[i]) & (meas$species %in% spgroups[[a]]),] <-
           temp
    }
  }
  #Now extract all specimens from this slide
  temp <- meas[(as.numeric(meas$age) == slides[i]),]
  #Calculate the average silicification for this slide
  meansipct[i] <- mean(temp$sipct)
  #Calculate the variance silicification for this slide
  varsipct[i] <- var(temp$sipct)
  #Keep track of sample size
  nsamp[i] <- dim(temp)[1]
}

#Reorder the mean values so that they plot in the right order
age <- slides
simean <- meansipct
sivar <- varsipct
TSdata <- data.frame(simean,sivar,nsamp,age)
TSdata <- TSdata[order(TSdata$age),]

didymoTS <- as.paleoTS(mm=TSdata$simean,vv=TSdata$sivar,nn=TSdata$nsamp,tt=TSdata$age,
     oldest="last")
didymosi <- simean
didymoage <- age

#Model fitting exercise (following Hunt 2006):
fit3models(didymoTS)

#Alrighty. Now let's go for Centrobotrys.
#Again, get the data from RadData.sqlite first
meas <- dbGetQuery(con, "SELECT holes.hole_id, holes.latitude, holes.longitude, slides.age,
      individuals.species, individuals.indiv_id, measurements.meas_type, measurements.meas_
     value FROM holes INNER JOIN slides ON holes.hole_id=slides.hole_id  INNER JOIN
     individuals ON slides.slide_id=individuals.slide_id INNER JOIN measurements ON
     measurements.indiv_id=individuals.indiv_id WHERE individuals.species LIKE '
     Centrobotrys%' ORDER BY individuals.indiv_id")

#Now, let's melt-cast this so that each row is a specimen so we can get silica %ages
names(meas) <- c("hole_id","latitude","longitude","age","species","indiv_id","variable","
     value")
meas <- cast(data=meas)

#These are the slide ages we have
slides <- unique(meas$age)
#Set up a variable to hold the average silicification for each slide
meansipct <- slides
#Set up a variable to hold the variance in silicification for each slide
varsipct <- slides
#And another one to keep track of sample size in each slide
nsamp <- slides

#In order to calculate silicifications, we need the geometric functions
source("RadShapeFunctions.R")

#The Centro species, grouped by geometric model
```

Code G.3: AnalyzeRadData.R (continued)

```
spgroups <- list("Centrobotrys gravida","Centrobotrys petrushevskayae","Centrobotrys thermophila")

meas$sipct <- NA

#Loop through each slide
for(i in 1:length(slides))
{
  #Also, loop through each set of species (grouped by geometric model)
  for(a in 1:length(spgroups))
  {
    #Extract the individuals from the current slide and current species set
    temp <- meas[(as.numeric(meas$age) == slides[i]) & (meas$species %in% spgroups[[a]]),]
    #If the extract isn't empty (i.e. if there are matching specimens)
    if(dim(temp)[1] > 0)
    {
      #Replace NAs with average values for that slide and current species set
      for(j in 7:19)
      {
        #Only do it if there are non-NA values in that column
        if(sum(is.na(temp[,j])) < length(temp[,j]))
        {
          temp[is.na(temp[,j]),j] <- mean(temp[,j],na.rm=TRUE)
        }
      }
      cat(a)
      #Now calculate the silicification percentages
      if(a == 1){temp$sipct <- getCgravSi(temp)}
      if(a == 2){temp$sipct <- getCpetSi(temp)}
      if(a == 3){temp$sipct <- getCthermSi(temp)}
      #Now copy that extract back into meas
      meas[(as.numeric(meas$age) == slides[i]) & (meas$species %in% spgroups[[a]]),] <- temp
    }
  }
  #Now extract all specimens from this slide
  temp <- meas[(as.numeric(meas$age) == slides[i]),]
  #Calculate the average silicification for this slide
  meansipct[i] <- mean(temp$sipct)
  #Calculate the variance silicification for this slide
  varsipct[i] <- var(temp$sipct)
  #Keep track of sample size
  nsamp[i] <- dim(temp)[1]
}

#Reorder the mean values so that they plot in the right order
age <- slides
simean <- meansipct
sivar <- varsipct
TSdata <- data.frame(simean,sivar,nsamp,age)
TSdata <- TSdata[order(TSdata$age),]

centroTS <- as.paleoTS(mm=TSdata$simean,vv=TSdata$sivar,nn=TSdata$nsamp,tt=TSdata$age, oldest="last")
centrosi <- simean
centroage <- TSdata$age

#Model fitting exercise (following Hunt 2006):
```

Code G.3: AnalyzeRadData.R (continued)

```
fit3models(centroTS)

#Now plot it up
#Get the fonts ready
file.exists <- function( fname ) length(Sys.glob(fname))>0
absolute.path.to.font.files <-  "/Users/bkotrc/font/";
bera.names <- c("gillsans","gillsansbold","gillsansitalic","gillsansbolditalic");
## if you do not have the correct font types
for (i in 1:length(bera.names)) {
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".afm", sep="")) )
  stopifnot( file.exists(paste(absolute.path.to.font.files,
                               bera.names[i], ".otf", sep="")) )
}
gillsans <- Type1Font("GillSans", paste(absolute.path.to.font.files,
                                        bera.names, ".afm", sep=""))
pdfname <- "radcomparison.pdf"
pdf(file=pdfname, bg="white", width=(7.2/cm(1)), height=(7.2/cm(1)), pointsize=7, family=
     gillsans);

plot(pnas$age,pnas$si,pch=16,col="#000000FF",ylim=c(0,20),xlim=c(65,0),axes=FALSE,xlab='
     Geologic Time (Ma)',ylab='Mean Silicification %age',bty="n",main="",xaxs="i",xpd=TRUE,
     type="o")
axthck <- 0.3
berg95(line=axthck)
axis(1, lwd.ticks=axthck, lwd=axthck, at=seq(60,0,-10),xpd=TRUE)
axis(2, lwd=axthck)
points(stichoage,stichosi,pch=15,type="o",col="#66000080")
points(centroage,centrosi,pch=17,type="o",col="#00660080")
#Reorder didymo:
didymo <- data.frame(didymoage,didymosi)
didymo <- didymo[order(didymo$didymoage),]
points(didymo$didymoage,didymo$didymosi,pch=15,type="o",col="#00006680")
#Legend
legend(x=60,y=8,legend=c("All Radiolaria","Stichocorys","Centrobotrys","Didymocyrtis"),col=
     c("black","#66000080","#00660080","#00006680"),pch=c(16,15,17,15),box.lwd=axthck,lty="
     solid",pt.cex=1,cex=0.75)


#Close connection to pdf file
dev.off()
#Now embed font in that file
embedFonts(file=pdfname,outfile=pdfname, fontpaths="/Users/bkotrc/font")

###END GRAND PLOT#############

#Make sure to close the connection before you quit.
dbDisconnect(con);
```

G.4 R FUNCTIONS FOR CALCULATING RADIOLARIAN SILICIFICATION

Code G.4: RadShapeFunctions.R

```
#Code to calculate silicification volumes from
```

Code G.4: RadShapeFunctions.R (continued)

```
#radiolarian morphometric measurements
#Benjamin Kotrc, February 2013
#kotrc@fas.harvard.edu

###BASIC SHAPE VOLUME FUNCTIONS###

#Calculate the volume of a cylinder
#Takes: radius of cylinder, height of cylinder
#Returns: volume of cylinder
cylinder <- function(r,h)
{
  vol <- pi*(r^2)*h
  return(vol)
}

#Calculate the volume of a sphere
#Takes: radius of sphere
#Returns: volume of sphere
sphere <- function(r)
{
  vol <- (4/3)*pi*(r^3)
  return(vol)
}

#Calculate the volume of a spheroid
#Takes: equatorial radius of spheroid (a), polar radius (c)
#Returns: volume of spheroid
spheroid <- function(a,c)
{
  vol <- (4/3)*pi*(a^2)*c
  return(vol)
}

#Calculate the volume of a cone
#Takes: radius of base (a)
#   height of cone (h)
#Returns: volume of cone
cone <- function(a,h)
{
  vol <- (1/3)*h*pi*(a^2)
  return(vol)
}

#Calculate the volume of a conical frustum
#Takes: radius of base (a)
#   radius of top (b)
#   height of frustum (h)
#Returns: volume of conical frustum
confrust <- function(a,b,h)
{
  vol <- (1/3)*((a^2)+(a*b)+(b^2))*h*pi
  return(vol)
}

#Calculate the volume of a spherical cap
#Takes: radius of sphere (r), height of spherical cap (h)
#Returns: volume of spherical cap
sphericalcap <- function(r,h)
```

Code G.4: RadShapeFunctions.R (continued)

```
{
  vol <- (1/3)*pi*(h^2)*((3*r)-h)
  return(vol)
}

#Calculate the volume of a spheroidal cap
#Takes: eq. radius of spheroid (r), polar radius (c),
#height of cap (h)
#Returns: volume of spheroid cap
spheroidcap <- function(r,c,h)
{
  vol <- ((pi*(r^2)*(h^2))/(3*(c^2)))*((3*c)-h)
  return(vol)
}

#Calculate the volume enclosed by two overlapping spheres
#Takes: diameter of spheres (d), length of intersecting shape (l)
#Returns: volume enclosed by spheres
doublesphere <- function(l,d)
{
  #Volume of one sphere
  spherevol <- sphere(d/2)
  #Height of one spherical cap
  caph <- ((2*d) - l)/2
  #Volume of one spherical cap
  capvol <- sphericalcap(d/2,caph)
  #Total volume enclosed by two overlapping spheres
  vol <- 2*(spherevol-capvol)
  return(vol)
}


###SILICIFICATION FUNCTIONS###

#Stichocorys species
#Takes: the 11 measurements for this geometric model,
#length of horn, width of horn, etc, as a data frame
#row (in any order).
#Returns: silicification %age
getStichoSi <- function(meas)
{
  Lhorn <- meas$'25'
  Whorn <- meas$'11'
  Wceph <- meas$'12'
  Wabd <- meas$'16'
  Wbase <- meas$'17'
  Labd <- meas$'30'
  Lpostabd <- meas$'28'
  Pabd <- meas$'44'
  Ppabd <- meas$'45'
  Tabd <- meas$'50'
  Tpabd <- meas$'51'
  #Shell is an outer volume less an inner volume
  #minus pore volume
  #Silicification %age is the shell volume as a
  #proportion of the outer volume
  #Outer volume:
  #Horn (cone)
```

Code G.4: RadShapeFunctions.R (continued)

```
  horn <- cone((Whorn/2),Lhorn)
  #Cephalis (sphere)
  cepho <- sphere(Wceph/2)
  #Upper abdomen
  uppero <- confrust(Wceph/2,Wabd/2,Labd)
  #Lower abdomen/postabdomen
  lowero <- confrust(Wbase/2,Wabd/2,Lpostabd)
  #Total outer volume
  outervol <- cepho+uppero+lowero+horn
  #Inner volume:
  #Cephalis (sphere)
  cephi <- sphere((Wceph/2)-Tabd)
  #Upper postabdomen
  upperi <- confrust((Wceph/2)-Tabd,(Wabd/2)-Tabd,Labd)
  #Lower abdomen/postabdomen
  loweri <- confrust((Wbase/2)-Tpabd,(Wabd/2)-Tpabd,Lpostabd)
  #Silicified volume:
  #Cephalis
  ceph <- cepho-cephi
  #Upper abdomen
  upper <- (uppero-upperi)*(1-Pabd)
  #Lower abdomen/postabdomen
  lower <- (lowero-loweri)*(1-Ppabd)
  #Total silicified volume
  sivol <- ceph+upper+lower+horn
  #Silicification percentage
  sipct <- (sivol/outervol)*100
  return(sipct)
}

#Didymocyrtis tetrathalamus
#Takes: 14 measurements for this geometric model,
#width of polar caps, width of cortical shell, etc,
#as a data frame row (in any order)
#Returns: silicification %age
getDtetraSi <- function(meas)
{
  Wcaps <- meas$'9'
  Wcshell <- meas$'4'
  Lcaps <- meas$'23'
  Lcshell <- meas$'20'
  Wimshell <- meas$'7'
  Womshell <- meas$'8'
  Pcaps <- meas$'41'
  Pcshell <- meas$'38'
  Tcaps <- meas$'48'
  Tcshell <- meas$'46'
  Wveil <- meas$'10'
  Lveil <- meas$'24'
  Pveil <- meas$'40'
  Tveil <- meas$'47'
  #Because so few specimens had a preserved veil,
  #and because including it would bias test volumes for
  #preservation affecting % silicification calculated,
  #veils were ignored in this version of the software
  #Shell is an outer volume less an inner volume,
  #minus pore volume, plus the medullary shells
  #Silicification %age is the shell volume as a
```

Code G.4: RadShapeFunctions.R (continued)

```
  #proportion of the outer volume
  #Outer volume:
  #Caps (two half-spheroids, one on either end of the cortical shell,
  #adds up to one spheroid)
  capo <- spheroid(Wcaps/2,Lcaps)
  #Double-sphere cortical shell
  cshello <- doublesphere(Lcshell,Wcshell)
  #Total outer volume:
  outervol <- capo + cshello
  #Inner volume:
  #Caps (two half-spheroids, one on either end of the cortical shell,
  #adds up to one spheroid)
  capi <- spheroid((Wcaps/2)-Tcaps,Lcaps-Tcaps)
  #Double-sphere cortical shell
  cshelli <- doublesphere(Lcshell-Tcshell,Wcshell-Tcshell)
  #Medullary shells
  outermshell <- (sphere(Womshell/2)-sphere((Womshell/2)-Tcshell))*(1-Pcshell)
  innermshell <- (sphere(Wimshell/2)-sphere((Wimshell/2)-Tcshell))*(1-Pcshell)
  #Silicified volume
  cap <- (capo-capi)*(1-Pcaps)
  cshell <- (cshello-cshelli)*(1-Pcshell)
  #Total silicified volume
  sivol <- cap+cshell+outermshell+innermshell
  #Silicification percentage
  sipct <- (sivol/outervol)*100
  return(sipct)
}

#Didymocyrtis prismatica
#Takes: 9 measurements for this geometric model,
#width of spongy columns, length of spongy columns, etc,
#as a data frame row (in any order)
#Optional boolean argument includearms determines whether
#spongy columns are included in the geometric model
#Returns: silicification %age
getDprisSi <- function(meas,includearms=TRUE)
{
  Wspdist <- meas$'1'
  Wspprox <- meas$'2'
  Wcshell <- meas$'4'
  Wmshell <- meas$'6'
  Lsp <- meas$'22'
  Lcshell <- meas$'20'
  Psp <- meas$'39'
  Pcshell <- meas$'38'
  Tcshell <- meas$'46'
  #Pore area on spongy columns proved unfeasible to measure,
  #so estimate it to be a (volumetric) porosity of 50%
  Psp <- 0.5
  #Because so few specimens had a preserved veil,
  #and because including it would bias test volumes for
  #preservation affecting % silicification calculated,
  #veils were ignored in this version of the software
  #Shell is an outer volume less an inner volume,
  #minus pore volume, plus the medullary shell
  #Silicification %age is the shell volume as a
  #proportion of the outer volume
  #Spongy columns
```

Code G.4: RadShapeFunctions.R (continued)

```
  spcolo <- confrust(Wspprox/2,Wspdist/2,Lsp)
  spcol <- spcolo*(1-Psp)
  #Outer volume:
  #Spheroid cortical shell
  cshello <- spheroid(Wcshell/2,Lcshell/2)
  #Total outer volume:
  outervol <- cshello
  if(includearms==TRUE)
  {
    outervol <- outervol+spcolo
  }
  #Inner volume:
  #Spheroid cortical shell
  cshelli <- spheroid((Wcshell/2)-Tcshell,(Lcshell/2)-Tcshell)
  #Medullary shell
  mshell <- (sphere(Wmshell/2)-sphere((Wmshell/2)-Tcshell))*(1-Pcshell)
  #Silicified volume
  cshell <- (cshello-cshelli)*(1-Pcshell)
  #Total silicified volume
  sivol <- cshell+mshell
  if(includearms==TRUE)
  {
    sivol <- sivol+spcol
  }
  #Silicification percentage
  sipct <- (sivol/outervol)*100
  return(sipct)
}

#Didymocyrtis penultima, antepenultima, avita, and laticonus
#Takes: 18 measurements for this geometric model,
#width of spongy columns, length of spongy columns, etc,
#as a data frame row (in any order)
#Optional boolean argument includearms determines whether
#spongy columns are included in the geometric model
#Returns: silicification %age
getDpenSi <- function(meas,includearms=TRUE)
{
  Wspdist <- meas$'1'
  Wspprox <- meas$'2'
  Wcaps <- meas$'9'
  Wcshell <- meas$'4'
  Wimshell <- meas$'7'
  Womshell <- meas$'8'
  Lsp <- meas$'22'
  Lcaps <- meas$'23'
  Lcshell <- meas$'20'
  Lplicae <- meas$'53'
  Psp <- meas$'39'
  Pcshell <- meas$'38'
  Tcshell <- meas$'46'
  Wveil <- meas$'10'
  Lveil <- meas$'24'
  Pveil <- meas$'40'
  Tveil <- meas$'47'
  Nplicae <- meas$'54'
  #Because so few specimens had a preserved veil,
  #and because including it would bias test volumes for
```

Code G.4: RadShapeFunctions.R (continued)

```
  #preservation affecting % silicification calculated,
  #veils were ignored in this version of the software
  #Pore area on spongy columns proved unfeasible to measure,
  #so estimate it to be a (volumetric) porosity of 50%
  Psp <- 0.5
  #Number of plicae did not get recorded due to a bug in the
  #RadDataInterface.R code, so for consistency assume there
  #are 4 plicae consistently
  Nplicae <- 4
  #Shell is an outer volume less an inner volume,
  #minus pore volume, plus the medullary shell
  #Silicification %age is the shell volume as a
  #proportion of the outer volume
  #Spongy columns
  spcolo <- confrust(Wspprox/2,Wspdist/2,Lsp)
  spcol <- spcolo*(1-Psp)
  #Plicae; modeled here as half-cylinders of radius Lplicae
  #with a thickness of Tcshell (think coin broken in half)
  plicae <- (cylinder(Lplicae,Tcshell)/2)*Nplicae
  #Outer volume:
  #Caps (two half-spheroids, one on either end of the cortical shell,
  #adds up to one spheroid)
  capo <- spheroid(Wcaps/2,Lcaps)
  #Double-sphere cortical shell
  cshello <- doublesphere(Lcshell,Wcshell)
  #Total outer volume:
  outervol <- capo + cshello + plicae
  if(includearms==TRUE)
  {
    outervol <- outervol+spcolo
  }
  #Inner volume:
  #Caps (two half-spheroids, one on either end of the cortical shell,
  #adds up to one spheroid)
  Tcaps <- Tcshell
  capi <- spheroid((Wcaps/2)-Tcaps,Lcaps-Tcaps)
  #Double-sphere cortical shell
  cshelli <- doublesphere(Lcshell-Tcshell,Wcshell-Tcshell)
  #Medullary shells
  outermshell <- (sphere(Womshell/2)-sphere((Womshell/2)-Tcshell))*(1-Pcshell)
  innermshell <- (sphere(Wimshell/2)-sphere((Wimshell/2)-Tcshell))*(1-Pcshell)
  #Silicified volume
  cap <- (capo-capi)*(1-Pcshell)
  cshell <- (cshello-cshelli)*(1-Pcshell)
  #Total silicified volume
  sivol <- cap+cshell+outermshell+innermshell+plicae
  if(includearms==TRUE)
  {
    sivol <- sivol+spcol
  }
  #Silicification percentage
  sipct <- (sivol/outervol)*100
  return(sipct)
}

#Didymocyrtis violina and mammifera
#Takes: 12 measurements for this geometric model,
#width of spongy columns, length of spongy columns, etc,
```

Code G.4: RadShapeFunctions.R (continued)

```
#as a data frame row (in any order)
#Optional boolean argument includearms determines whether
#spongy columns are included in the geometric model
#Returns: silicification %age
getDvioSi <- function(meas,includearms=TRUE)
{
  Wspdist <- meas$'1'
  Wspprox <- meas$'2'
  Wcshell <- meas$'4'
  Wimshell <- meas$'7'
  Womshell <- meas$'8'
  Lsp <- meas$'22'
  Lcshell <- meas$'20'
  Lplicae <- meas$'53'
  Psp <- meas$'39'
  Pcshell <- meas$'38'
  Tcshell <- meas$'46'
  Nplicae <- meas$'54'
  #Pore area on spongy columns proved unfeasible to measure,
  #so estimate it to be a (volumetric) porosity of 50%
  Psp <- 0.5
  #Number of plicae did not get recorded due to a bug in the
  #RadDataInterface.R code, so for consistency assume there
  #are 4 plicae consistently
  Nplicae <- 4
  #Shell is an outer volume less an inner volume,
  #minus pore volume, plus the medullary shell
  #Silicification %age is the shell volume as a
  #proportion of the outer volume
  #Spongy columns
  spcolo <- confrust(Wspprox/2,Wspdist/2,Lsp)
  spcol <- spcolo*(1-Psp)
  #Plicae; modeled here as half-cylinders of radius Lplicae
  #with a thickness of Tcshell (think coin broken in half)
  plicae <- (cylinder(Lplicae,Tcshell)/2)*Nplicae
  #Outer volume:
  #Double-sphere cortical shell
  cshello <- doublesphere(Lcshell,Wcshell)
  #Total outer volume:
  outervol <- cshello + plicae
  if(includearms==TRUE)
  {
    outervol <- outervol+spcolo
  }
  #Inner volume:
  #Double-sphere cortical shell
  cshelli <- doublesphere(Lcshell-Tcshell,Wcshell-Tcshell)
  #Medullary shells
  outermshell <- (sphere(Womshell/2)-sphere((Womshell/2)-Tcshell))*(1-Pcshell)
  innermshell <- (sphere(Wimshell/2)-sphere((Wimshell/2)-Tcshell))*(1-Pcshell)
  #Silicified volume
  cshell <- (cshello-cshelli)*(1-Pcshell)
  #Total silicified volume
  sivol <- cshell+outermshell+innermshell+plicae
  if(includearms==TRUE)
  {
    sivol <- sivol+spcol
  }
```

Code G.4: RadShapeFunctions.R (continued)

```
  #Silicification percentage
  sipct <- (sivol/outervol)*100
  return(sipct)
}

#Centrobotrys gravida
#Takes: 8 measurements for this geometric model,
#length of thorax, width of thorax, etc,
#as a data frame row (in any order)
#Returns: silicification %age
getCgravSi <- function(meas)
{
  Ltot <- meas$'34'
  Lthorax <- meas$'35'
  Ltop <- meas$'36'
  Wtop <- meas$'18'
  Wthorax <- meas$'13'
  Wceph <- meas$'12'
  Pthorax <- meas$'43'
  Tthorax <- meas$'49'
  #Shell is an outer volume less an inner volume,
  #minus pore volume, plus the eucephalic shell
  #Silicification %age is the shell volume as a
  #proportion of the outer volume
  #Outer volume:
  #Post/antecephalic chamber (spheroid less a spheroid cap)
  #Length of entire spheroid
  Lpacepho <- Ltop+Lthorax-Ltot+Ltop
  pacepho <- spheroid(Wtop/2,Lpacepho/2)
  #Subtract out the cap that is missing where the post/antecephalic
  #chamber meets the thorax; approximate this length (not measured directly)
  pacepho <- pacepho - spheroidcap(Wtop/2,Lpacepho/2,(Ltop+Lthorax-Ltot))
  #Thorax
  thoraxo <- spheroid(Wthorax/2,Lthorax/2)
  #Total outer volume:
  outervol <- pacepho+thoraxo
  #Inner volume:
  #Post/antecephalic chamber (spheroid less a spheroid cap)
  pacephi <- spheroid((Wtop/2)-Tthorax,(Lpacepho/2)-Tthorax)
  #Subtract out the cap that is missing where the post/antecephalic
  #chamber meets the thorax; approximate this length (not measured directly)
  pacephi <- pacephi - spheroidcap((Wtop/2)-Tthorax,(Lpacepho/2)-Tthorax,(Ltop+Lthorax-Ltot
       ))
  #Thorax
  thoraxi <- spheroid((Wthorax/2)-Tthorax,(Lthorax/2)-Tthorax)
  #Eucephalic lobe
  euceph <- sphere(Wceph/2)-sphere((Wceph/2)-Tthorax)
  #Silicified volume
  paceph <- (pacepho-pacephi)*(1-Pthorax)
  thorax <- (thoraxo-thoraxi)*(1-Pthorax)
  #Total silicified volume
  sivol <- paceph+thorax+euceph
  #Silicification percentage
  sipct <- (sivol/outervol)*100
  return(sipct)
}

#Centrobotrys petrushevskayae
```

Code G.4: RadShapeFunctions.R (continued)

```
#Takes: 8 measurements for this geometric model,
#length of thorax, width of thorax, etc,
#as a data frame row (in any order)
#Returns: silicification %age
getCpetSi <- function(meas)
{
  Ltot <- meas$'34'
  Lthorax <- meas$'35'
  Ltop <- meas$'36'
  Wtop <- meas$'18'
  Wthorax <- meas$'14'
  Wceph <- meas$'12'
  Pthorax <- meas$'43'
  Tthorax <- meas$'49'
  #Shell is an outer volume less an inner volume,
  #minus pore volume, plus the eucephalic shell
  #Silicification %age is the shell volume as a
  #proportion of the outer volume
  #Outer volume:
  #Post/antecephalic chamber (spheroid less a spheroid cap)
  #Length of entire spheroid
  Lpacepho <- Ltop+Lthorax-Ltot+Ltop
  pacepho <- spheroid(Wtop/2,Lpacepho/2)
  #Subtract out the cap that is missing where the post/antecephalic
  #chamber meets the thorax; approximate this length (not measured directly)
  pacepho <- pacepho - spheroidcap(Wtop/2,Lpacepho/2,(Ltop+Lthorax-Ltot))
  #Thorax
  thoraxo <- cylinder(Wthorax/2,Lthorax)
  #Total outer volume:
  outervol <- pacepho+thoraxo
  #Inner volume:
  #Post/antecephalic chamber (spheroid less a spheroid cap)
  pacephi <- spheroid((Wtop/2)-Tthorax,(Lpacepho/2)-Tthorax)
  #Subtract out the cap that is missing where the post/antecephalic
  #chamber meets the thorax; approximate this length (not measured directly)
  pacephi <- pacephi - spheroidcap((Wtop/2)-Tthorax,(Lpacepho/2)-Tthorax,(Ltop+Lthorax-Ltot
      ))
  #Thorax
  thoraxi <- cylinder((Wthorax/2)-Tthorax,Lthorax-Tthorax)
  #Eucephalic lobe
  euceph <- sphere(Wceph/2)-sphere((Wceph/2)-Tthorax)
  #Silicified volume
  paceph <- (pacepho-pacephi)*(1-Pthorax)
  thorax <- (thoraxo-thoraxi)*(1-Pthorax)
  #Total silicified volume
  sivol <- paceph+thorax+euceph
  #Silicification percentage
  sipct <- (sivol/outervol)*100
  return(sipct)
}

#Centrobotrys thermophila
#Takes: 8 measurements for this geometric model,
#length of thorax, width of thorax, etc,
#as a data frame row (in any order)
#Returns: silicification %age
getCthermSi <- function(meas)
{
```

Code G.4: RadShapeFunctions.R (continued)

```
  Ltop <- meas$'37'
  Lthorax <- meas$'35'
  Wtop <- meas$'52'
  Wceph <- meas$'12'
  Wmid <- meas$'19'
  Wbase <- meas$'17'
  Pthorax <- meas$'43'
  Tthorax <- meas$'49'
  #Shell is an outer volume less an inner volume,
  #minus pore volume, plus the eucephalic shell
  #Silicification %age is the shell volume as a
  #proportion of the outer volume
  #Outer volume:
  #Post/antecephalic chamber (conical frustum)
  pacepho <- confrust(Wmid/2,Wtop/2,Ltop)
  #Thorax
  thoraxo <- confrust(Wbase/2,Wmid/2,Lthorax)
  #Total outer volume:
  outervol <- pacepho+thoraxo
  #Inner volume:
  #Post/antecephalic chamber (conical frustum)
  pacephi <- confrust((Wmid/2)-Tthorax,(Wtop/2)-Tthorax,Ltop-Tthorax)
  #Thorax
  thoraxi <- confrust((Wbase/2)-Tthorax,(Wmid/2)-Tthorax,Lthorax)
  #Eucephalic lobe
  euceph <- sphere(Wceph/2)-sphere((Wceph/2)-Tthorax)
  #Silicified volume
  paceph <- (pacepho-pacephi)*(1-Pthorax)
  thorax <- (thoraxo-thoraxi)*(1-Pthorax)
  #Total silicified volume
  sivol <- paceph+thorax+euceph
  #Silicification percentage
  sipct <- (sivol/outervol)*100
  return(sipct)
}
```

G.5 ImageJ macros

Code G.5: ImageJmacros.txt

```
//Macros to collect radiolarian morphometric measurements for acquisition
//to the RadData database, via its R interface.
//Written by Ben Kotrc, Mar 1st, 2011
//Append to "StartupMacros.txt" file for automatic loading
//See file "How to set up ImageJ.txt" for detailed instructions

//Declare global variables
//Measurement number (in reference to R interface image)
var counter = 0;
//Microscope magnification (e.g. 16x, 40x, etc)
var objective = "??x";

//Macro to write all measurements of type "Length" taken since the last
//file write operation, with the current image file name, to the pipe file
```

Code G.5: ImageJmacros.txt (continued)

```
macro ”Save linear measurements to file [1]” {
//Get name of currently open image file
curfname=getInfo(”image.filename”);
//Loop through each row in the Results table
for (i=counter; i<nResults; i++){
  //Extract the ith row from the ”Length” column of the Results table
  measurement = getResult(”Length”, i);
  //Concatenate row of data to be appended to pipe file (data + filename)
  insertion = d2s(measurement,2) + ”\t” + curfname;
  //Append the ith measurement to the Results table
  File.append(insertion, ”/Users/bkotrc/Dropbox/Harvard/By-Lineage\ Rads/RadData\ Database/
       pipefilename.txt”);
  //Increment the measurement counter
  counter++;
}
//Display in log window how many measurements have been written to file
print(”\\Clear”);
print(”Scale set to ” + objective + ”.”);
print(d2s(counter,0) + ” measurements recorded to file.”);
//Set the image magnification in the metadata of the image file
setMetadata(”Info”, objective);
}

//Macro to calculate the pore area proportion from the measurements added
//to the results table since the last file write operation, then write the
//number to file with the current image file name
macro ”Save area measurement to file [2]” {
//Get name of currently open image file
curfname=getInfo(”image.filename”);
//Find the sum of all but the first row in the Results table
sum=0;
for (i=counter+1; i<nResults; i++){
  sum = sum + getResult(”Area”,i);
}
//Find the pore area proportion
measurement = sum/getResult(”Area”,counter);
//Concatenate row of data to be appended to pipe file (data + filename)
insertion = d2s(measurement,4) + ”\t” + curfname;
//Append the result to file
File.append(insertion, ”/Users/bkotrc/Dropbox/Harvard/By-Lineage\ Rads/RadData\ Database/
     pipefilename.txt”);
//Increment the measurement counter
counter++;
//Display in log window how many measurements have been written to file
print(”\\Clear”);
print(”Scale set to ” + objective + ”.”);
print(d2s(counter,0) + ” measurements recorded to file.”);
//Tidy up the Results table
IJ.deleteRows(counter,nResults-1);
setResult(”Area”,counter-1,measurement);
//Set the image magnification in the metadata of the image file
setMetadata(”Info”, objective);
}

//Macro to insert a NA-valued (-1) measurement to the pipe file, and update
//the ImageJ results table accordingly
macro ”Add zero measurement to file [3]” {
measurement = -1;
```

Code G.5: ImageJmacros.txt (continued)

```
curfname = ”No file”;
//Concatenate row of data to be appended to pipe file (data + filename)
insertion = d2s(measurement,4) + ”\t” + curfname;
//Append the result to file
File.append(insertion, ”/Users/bkotrc/Dropbox/Harvard/By-Lineage\ Rads/RadData\ Database/
     pipefilename.txt”);
//Increment the measurement counter
counter++;
//In case this is the first measurement, make a dummy measurement
//to open the results table
run(”Measure”);
//Update Results table to reflect added zero ”measurement”
setResult(”Area”, counter-1, -1);
//Display in log window how many measurements have been written to file
print(”\\Clear”);
print(”Zero added. Scale set to ” + objective + ”.”);
print(d2s(counter,0) + ” measurements recorded to file.”);
}

//Macro to reset the tally of the number of measurements written to file,
//as displayed in the log window, to zero (use each time a new individual
//is started)
macro ”Reset measurement counter [4]” {
//Clear the Results table
selectWindow(”Results”);
run(”Close”);
//Reset counter
counter = 0;
//Display in log window
print(”\\Clear”);
print(”Scale set to ” + objective + ”.”);
print(d2s(counter,0) + ” measurements recorded to file.”);
}

//Macro to set the appropriate scale for different microscope objectives
macro ”Set_scale” {
Dialog.create(”Set scale”);
Dialog.addChoice(”Objective:”, newArray(”4x”, ”10x”, ”25x”, ”40x”, ”60x”));
Dialog.show();
objective = Dialog.getChoice();
if (objective == ”4x”) {
  run(”Set Scale...”, ”distance=4439 known=1700 pixel=1 unit=µm”);
} else if (objective == ”10x”){
  run(”Set Scale...”, ”distance=3895 known=600 pixel=1 unit=µm”);
} else if (objective == ”25x”){
  run(”Set Scale...”, ”distance=3469 known=210 pixel=1 unit=µm”);
} else if (objective == ”40x”){
  run(”Set Scale...”, ”distance=4259 known=160 pixel=1 unit=µm”);
} else if (objective == ”60x”){
  run(”Set Scale...”, ”distance=3937 known=100 pixel=1 unit=µm”);
}
//Print new magnification in log window, along with # of measurements taken
print(”\\Clear”);
print(”Scale set to ” + objective + ”.”);
print(d2s(counter,0) + ” measurements recorded to file.”);
//Set the image magnification in the metadata of the image file
setMetadata(”Info”, objective);
//Append magnification at end of filename
```

Code G.5: ImageJmacros.txt (continued)

```
curfname=getInfo("image.filename");
index=lastIndexOf(curfname,".");
curfname=substring(curfname, 0, index);
rename(curfname+"_"+objective+".JPG");
//Save changes to name (this will create a second copy of the file)
saveAs("Jpeg", getDirectory("image") + getTitle());
}
```

Colophon

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Arno Pro, designed by Robert Slimbach in the style of book types from the Aldine Press in Venice, and issued by Adobe in 2007. The layout of this PhD thesis was based on a template released under the permissive MIT (X11) license that can be found online at github.com/suchow/, and on a modified version of that template that can be found at github.com/aleifer/.